Policy Research Working Paper 8261

Teaching with the Test

Experimental Evidence on Diagnostic Feedback and Capacity Building for Public Schools in Argentina

Rafael de Hoyos
Alejandro J. Ganimian
Peter A. Holland

WORLD BANK GROUP
Education Global Practice Group
November 2017

Policy Research Working Paper 8261

Abstract

Despite the recent growth in the number of large-scale student assessments, there is little evidence on their potential to inform improvements in school management and classroom instruction in developing countries. This study conducted an experiment in the Province of La Rioja Argentina, that randomly assigned 105 public primary schools to: (a) a "diagnostic feedback" group in which standardized tests were administered in math and reading comprehension at baseline and two follow-ups and the results were made available to the schools through user-friendly reports; (b) a "capacity-building" group for which schools were provided with the reports and also workshops and school visits for supervisors, principals, and teachers; or (c) a control group, in which the tests were administered only at the second follow-up. After two years, diagnostic feedback schools outperformed control schools by .34 and .36 standard deviations (SD) in third grade math and reading, and by .28 and .38 SD in fifth grade math and reading. The principals at these schools were more likely to report using assessment results for management decisions, and students were more likely to report that their teachers engaged in more instructional activities and improved their interactions with them. Capacity-building schools saw more limited impacts due to lower achievement at baseline, low take up, and little value-added of workshops and visits. However, in most cases the results cannot discard the possibility that both interventions had the same impact.

This paper is a product of the Education Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at rdehoyos@worldbank.org.

Produced by the Research Support Team

Teaching *with* the Test: Experimental Evidence on Diagnostic Feedback and Capacity Building for Public Schools in Argentina[*]

Rafael de Hoyos[†]
Alejandro J. Ganimian[‡]
Peter A. Holland[§]

JEL codes: C93, I21, I22, I25.

Keywords: Diagnostic feedback, capacity building, large-scale assessments, Argentina.

[*] We gratefully acknowledge the funding provided by the World Bank for this study. We thank Rita Abdala, Walter Flores, Silvia Romero, and Juan Vegas at the Ministry of Education of La Rioja for making this study possible. We also thank Eduardo Cascallar, Jorge Fasce, and Gustavo Iaies for designing the standardized tests and training workshops in this study. María Cortelezzi, Alvina Erman, Bárbara Funtowicz, Andrés Felipe Pérez, Melissa Rofman, María José Vargas, and Guillermo Toral were responsible for project management. We thank Lawrence Aber, Felipe Barrera-Osorio, Elise Cappella, Joseph Cimpian, Sean Corcoran, Andreas de Barros, Emmerich Davies, Andrew Ho, James Kim, Daniel Koretz, Isaac Mbiti, Richard Murnane, Jeffrey Puryear, Jonah Rockoff, Daniel Santín Gonzalez, Jennifer Steele, Alice Wuermli, Hirokazu Yoshikawa, and seminar participants at APPAM, the *Fundación Ramón Areces*, HGSE, NYU, and the World Bank country office in Argentina for useful comments that informed this draft. The usual disclaimers apply.

[†] Lead Economist, Education, The World Bank. E-mail: `rdehoyos@worldbank.org`.

[‡] Assistant Professor of Applied Psychology and Economics, New York University Steinhardt School of Culture, Education, and Human Development. E-mail: `alejandro.ganimian@nyu.edu`.

[§] Program Leader, Education, The World Bank. E-mail: `pholland@worldbank.org`.

1 Introduction

Over the past two decades, the international community has progressively shifted its attention from expanding access to schooling to ensuring all children achieve basic standards of learning. In the Millennium Development Goals, adopted by the United Nations General Assembly in 2000, 191 countries pledged to ensure that "by 2015, children everywhere, boys and girls alike will be able to complete a full course of primary schooling" (UNGA 2000). In the Sustainable Development Goals, adopted in 2015, 194 countries set a new target: "by 2030... all girls and boys [should] complete free, equitable, and *quality* primary and secondary education learning to relevant and effective *learning outcomes*" (UNGA 2015, emphasis added). This shift was partly informed by research documenting that expansions in schooling did not translate into commensurate progress in learning (Pritchett 2013) and that attainment without achievement is unlikely to improve wages or economic growth (Hanushek and Woessmann 2007, 2010).

This impetus for ensuring minimum learning standards has led many school systems to administer large-scale student assessments and to participate in international assessments. According to one mapping effort, 85 national school systems have conducted 306 assessments of math, language, and science since 2004 (Cheng and Gale 2014). A similar effort found that 328 national and sub-national school systems have participated in 37 international assessments of the same subjects from 1963 to 2015; nearly half of them began participating since 1995 (Ganimian and Koretz 2017).[1] This exponential growth in large-scale assessments has been partly motivated by a belief in the potential of such assessments to help school systems identify learning gaps and inform reforms to remedy them. In its latest World Development Report, the World Bank recommends developing countries to "assess learning to make it a serious goal—measure and track learning better; use results to guide action" (World Bank 2017).

Yet, there seems to be a disconnect between the rhetoric about the intended formative use of large-scale assessments and the existing evidence on their impact in developing countries. Experimental and quasi-experimental evaluations have focused on whether these assessments can be used for accountability purposes (see, for example, Andrabi et al. 2017; Camargo et al. 2011; Mizala and Urquiola 2013). Some studies have explored whether classroom assessments can inform differentiated or scripted instruction (see, for example, Banerjee et al. 2011; Duflo et al. 2015; Piper and Korda 2011), but these tests differ considerably from the large-scale assessments that have rapidly grown in popularity in recent years. To our knowledge, there are only two impact evaluations of whether large-scale assessments can inform school management

[1] This figure is likely to increase further as the leading testing agencies develop assessments for low-income countries, which are reluctant to join existing global tests (e.g., the Organization for Economic Cooperation and Development's Program for International Student Assessment for Development and the International Association for the Evaluation of Educational Achievement's Literacy and Numeracy Assessment).

and/or classroom instruction and they reach conflicting conclusions (see de Hoyos et al. 2017; Muralidharan and Sundararaman 2010, discussed in detail below).[2]

This paper builds on existing evidence on the potential of large-scale assessments to inform improvements in school management and classroom instruction in developing countries. We randomly assigned 105 public primary schools in urban and semi-urban areas in the Province of La Rioja, Argentina to one of three groups: (a) a "diagnostic feedback" or T1 group, in which we administered standardized tests in math and reading comprehension at baseline and two follow-ups and made their results available to the schools through user-friendly reports; (b) a "capacity building" or T2 group, in which we also conducted professional development workshops and school visits; or (c) a control group, in which we administered the tests only at the second follow-up. We wanted to understand whether disseminating the assessments results was sufficient to prompt improvements in how schools were organized and how classes were taught, or whether dissemination needed to be complemented with support to distill the results for principals and teachers and to help them identify strategies to improve them.

After two years, T1 schools outperformed control schools by .34 and .36σ in third grade math and reading, and by .28 and .38σ in fifth grade math and reading, respectively. Student achievement improved in nearly all content and cognitive domains in both subjects. Consistent with these effects, principals at T1 schools were more likely than their control counterparts to report using assessment results to inform school management (e.g., evaluating teachers, making changes in the curriculum, or informing parents about school quality). Students at these schools were more prone than their control peers to report that their teachers engaged in more instructional activities (e.g., copying from the blackboard, explaining topics, and assigning and grading homework). They were also more likely to report positive student-teacher interactions (e.g., teachers being nice to them when they ask for help, explaining concepts in multiple ways, and checking that they understand the material).

In spite of being assigned to receive both diagnostic feedback *and* capacity-building activities, T2 schools only outperformed control schools in grade 5, by .21σ in math and .19σ in reading. This seems to be due to three main reasons. First, by chance, the schools that were randomly assigned to the T2 group were already performing considerably below those in the T1 group at baseline. Second, T2 schools participated in fewer workshops and school visits than expected. Third, each capacity-building activity (i.e., workshop or visit) had a positive but limited and statistically insignificant impact on achievement. Consistent with these effects, we found less clear evidence of impact mechanisms in T2 schools. Principals at these schools were more likely than their control counterparts to report using assessment results to inform school management, but students were no more likely to report changes in instruction. In nearly all

[2] This question is likely to become increasingly important as testing agencies have started selling school-level versions of their large-scale assessments, such as the OECD's PISA-based Test for Schools.

grades and subjects, we cannot discard the possibility that diagnostic feedback alone had the same effect when combined with capacity building.

The rest of the paper is structured as follows. Section 2 reviews prior research. Section 3 describes the context, intervention, sampling strategy, and randomization. Section 4 presents the data collected for this study. Section 5 discusses the empirical strategy. Section 6 reports the results. Section 7 discusses the implications for policy and further research.

2 Prior research

The bulk of prior experimental and quasi-experimental research on large-scale assessments in developing countries has focused on whether they can be used for accountability purposes, such as helping parents make better-informed choices about where to send their children to school (see, for example, Andrabi et al. 2017; Camargo et al. 2011; Mizala and Urquiola 2013). A few studies have explored whether classroom assessments can inform differentiated or scripted instruction (see, for example, Banerjee et al. 2011; Duflo et al. 2015; Piper and Korda 2011). These tests, however, are only used in a handful of countries and they differ considerably from the large-scale assessments that have rapidly grown in popularity in recent years.

To our knowledge, there have only been two impact evaluations of initiatives using large-scale assessments to inform school management and classroom instruction in developing countries: Muralidharan and Sundararaman (2010) in India and de Hoyos et al. (2017) in Mexico. Yet, these two studies reach seemingly conflicting conclusions.

Muralidharan and Sundararaman (2010) evaluated a program in the state of Andhra Pradesh, India in which 200 rural primary schools were randomly assigned to: (a) a treatment group of 100 schools, in which the Azim Premji Foundation (APF) administered standardized tests of math and language to children in grades 1 to 5 at the beginning of the 2005-2006 school year, provided these schools with student-, grade-, and school-level reports about a month later, and assessed these students again at the end of the school year; or (b) a comparison group of 100 schools, in which APF administered the same tests only at the end of the school year.[3] APF also visited treatment schools six times and control schools one time during the school year to collect data on teacher attendance and activity.[4] By the end of the school year, the authors found that treatment schools performed on par with control schools, even if teachers in treatment schools seemed to exert more effort during the unannounced visits.

[3] Treatment schools had been notified that they would be assessed at the end of the school year, but control schools only received a two-week notice to avoid prompting behavioral responses.

[4] The comparison group also included 200 schools in which APF conducted one unannounced visit during the school year to collect data on teacher attendance and activity, but did not administer standardized tests.

This study may give the impression that principals and teachers in developing countries are unwilling and/or unable to use assessment results to improve management and instruction.[5] Yet, there are two questions that it does not address. First, it evaluates the impact of diagnostic feedback over a school year (about nine months), but it is possible that principals and teachers needed more time to experiment and identify their optimal responses to improve learning outcomes. Second, the treatment group received two interventions that may have impacted schools in different ways: the reports may have encouraged principals and teachers to experiment with new practices, while the six unannounced visits may have led them to adopt practices that they believed were being tracked, crowding out experimentation.

de Hoyos et al. (2017) evaluated a program in the state of Colima, Mexico in which the 108 lowest-performing public primary schools on the national large-scale assessment received two types of interventions during the 2009-2010 and 2010-2011 school years. In 2009-2010, schools received: (a) a technical advisor from the state ministry of education who helped principals and teachers identify areas for improvement in the assessment and develop a school improvement plan and who visited them three times a month to monitor the plan's implementation; and (b) an online portal with the assessment's results on each item. In 2010-2011, schools received one or more of the following: (a) strengthening of school management, drawing on the experience of two national school management programs; (b) redefinition of the roles of and professional development for school supervisors and principals; and/or (c) professional development for teachers in the academic areas identified as needing support. The authors exploit the sudden introduction of the program and its eligibility cutoff to evaluate it using a difference-in-difference (DID) and a regression discontinuity (RDD), respectively. They find an effect of about $.12\sigma$ in math under both specifications (which is only statistically significant in the case of the DID) and an effect of $.07\sigma$ (DID) or $.2\sigma$ (RDD) in Spanish. Importantly, the improvement in test scores took place at the end of 2009-2010.

This study suggests that large-scale assessments may be used to improve management and instruction, but faces two limitations. First, it evaluates the effect of a bundled intervention with multiple (and different) components, so it is not possible to identify which components were responsible for the positive effects.[6] Second, as the authors acknowledge, each empirical strategy has its own shortcomings: in the DID, the parallel trends assumption is unlikely to be met, and in the RDD, several of the estimations lack sufficient statistical power.

[5] The authors conclude that "diagnostic feedback to teachers by itself may not be enough to improve student learning outcomes, especially in the absence of improved incentives to make effective use of the additional inputs" (Muralidharan and Sundararaman 2010, p. F189).

[6] The authors argue that the first phase of the program captures the effect of components related to accountability whereas the second captures that of components related to pedagogy. However, the first combines the effect of capacity building (from the technical advisor) with that of diagnostic feedback (from the online platform), while the second combines up to four different components. Additionally, the authors do not know which of the components in the second phase were implemented in each school.

It is not clear why these studies arrive at conflicting conclusions. It could be due to a number of differences between the two evaluations, including context, intervention, and/or study design. One potential explanation is that feedback and capacity building may have less scope for impact in lower-middle-income countries like India, where the binding constraint for improving learning outcomes is the *extensive* margin of principal and teacher effort (i.e., improving attendance), than in upper-middle-income countries like Mexico, where the binding constraint is the *intensive* margin of effort (i.e., improving productivity, conditional on attendance).[7] This interpretation is consistent with prior studies of teacher attendance and time-on-task (Abadzi 2007; Bruns and Luque 2014; Chaudhury et al. 2006; Kremer et al. 2005; Muralidharan et al. 2017; Sankar and Linden 2014; Stallings et al. 2014; World Bank 2016).

3 Experiment

3.1 Context

Schooling in Argentina is compulsory from the age of 4 until the end of secondary school. In 12 of the country's 24 provinces, including the Province of La Rioja, primary school runs from grades 1 to 7 and secondary school runs from grades 8 to 12 (DiNIECE 2013).[8] According to the latest official figures, the Argentine school system serves 11.1 million students: 1.7 million in pre-school, 4.5 million in primary, and 3.9 million in secondary school (DiNIEE 2015).

Argentina achieved near-universal access to primary education before most of Latin America: by the early 1990s, 95% of primary-school age children were enrolled in time, compared to 81% in the average country in the region. Argentina also has one of the highest primary school graduation rates in Latin America: by the late 2000s, 87% of primary school age children had graduated, compared to 76% in the average country in the region (Busso et al. 2013).

Yet, the relative performance of Argentina's primary school students in Latin America has deteriorated. In 1997, on the first regional assessment of reading, math, and science in primary school, Argentine third graders had ranked second in math, after their Cuban counterparts. In 2013, on the third regional assessment, they ranked seventh—on par with their peers in Peru and Ecuador, who had ranked near the bottom in 1997 and 2006 (Ganimian 2014).[9]

[7] We do not mean to imply that teacher absenteeism is not a problem in upper-middle-income countries. However, existing evidence indicates that countries like Ecuador and Peru have much lower absence rates (14 and 11%, respectively) than countries like India or Uganda (25 and 27%) (Chaudhury et al. 2006).

[8] In the other 12 provinces, primary runs from grades 1 to 6 and secondary from grades 7 to 12.

[9] The 1997 and 2006 assessments are not strictly comparable, but no other country participating in both assessments has changed its ranking so radically. Further, the deterioration in the relative standing of Argentine students is also seen in secondary (de Hoyos et al. 2015; Ganimian 2013).

Education policy in Argentina is shaped by both the national and sub-national (province) governments. According to the National Education Law of 2006, the provinces are responsible for pre-school, primary, and secondary education, and the federal government for higher education and for providing financial and technical assistance to the provinces. The national large-scale assessment is conducted by the Secretary of Educational Assessment at the National Ministry of Education and Sports, together with its counterparts in each province. Only a few provinces, including the City of Buenos Aires, also administer sub-national assessments.

Argentina is an interesting setting to evaluate the impact of using large-scale assessments for diagnostic feedback and capacity building for schools. Over the past two decades, the country has taken multiple steps that limited the generation, dissemination, and use of student achievement data: (a) it reduced the frequency of its national assessment from an annual basis (in 1999-2000), to a biennial basis (in 2002-2007), to a triennial basis (in 2008-2013); (b) it prohibited the public dissemination of any educational indicator at the level of the student, teacher, or school by law; and (c) in 2013, it discontinued the publication of the national assessment results at the province level, publishing them instead at the regional level (i.e., by groups of provinces) for the first time in history (Ganimian 2015). These policies stood in stark contrast with those of the rest of upper-middle-income countries in Latin America (e.g., Brazil, Colombia, Mexico, and Peru), which have technically robust and long-standing national LSAs and use them for multiple purposes (Ferrer 2006; Ferrer and Fiszbein 2015).

In 2015, a new government reversed several of the previous policies: (a) it published the abysmally low response rates of the latest national assessment (e.g., only 66% in grade 12, where the assessment was meant to be census-based) (Duro et al. 2016); (b) it adopted a new national assessment, to be administered annually, cover all students at the end of primary and secondary school (grades 7 and 12) and a sample of students halfway through each level (grades 3 and 8), and assess math, language, and natural and social sciences (SEE-MEDN 2016); and (c) it started distributing school-level reports of the results of the national assessment to all schools. Thus, the questions explored in this paper are not only of general interest to developing countries, but also of specific interest to Argentina.

We conducted our study in the Province of La Rioja for multiple reasons. First, it is one of the lowest-performing provinces in Argentina, so it stands to benefit considerably from policies that improve learning. The latest national assessment found that 41% of sixth graders in La Rioja performed at the two lowest levels in math and 53% in reading (SEE-MEDN 2017). Second, it is one of the smallest sub-national school systems in the country, which makes it easier to implement a school quality assurance mechanism. With 377 primary schools and 41,571 students at that level, it is the seventh-smallest system in terms of number of schools and the fourth-smallest in terms of students (DiNIEE 2015). Third, in 2013, it was one of the few provinces with the political will to experiment with a sub-national assessment. The assessment was endorsed by both the governor and the minister of education of the province.

3.2 Sample

The sampling frame for the study included all 126 public primary schools located in urban and semi-urban areas of La Rioja. We selected this frame as follows. First, out of the 394 primary schools in the province, we excluded the 29 private schools because we were interested in the potential of our interventions to improve the performance of public schools. Then, out of the 365 public primary schools, we excluded the 239 schools in rural areas because they are spread across the province, which would have limited our ability to monitor the implementation of the intervention. It is worth noting, however, that while rural schools account for a large share of the total number of public schools in La Rioja (65% of the total), they serve a small share of the students (less than 10%). The sample of 105 urban and semi-urban public primary schools was drawn randomly from the 126 schools and stratified by enrollment terciles.

In-sample schools differ from out-of-sample schools. First, in-sample schools have more students than all out-of-sample schools (i.e., all 239 rural schools, as well as the 33 urban and 8 semi-urban schools that were not selected for the study) (Table B.1 in Appendix B). This difference is driven by rural schools, which as we already mentioned, are much smaller than urban and semi-urban schools. Yet, in-sample schools also have more students than urban and semi-urban out-of-sample schools (i.e., only the 33 urban and 8 semi-urban schools that were not selected for the study). Second, in-sample schools are also more likely to be what the province calls “category 1” schools (be urban or semi-urban *and* have both a principal and a vice-principal) or “category 2” schools (be urban or semi-urban *and* have a principal) than all out-of-sample schools. Yet, we cannot determine whether this is driven by the fact that we selected urban and semi-urban schools or due to differences in the management structure of these schools because the province does not record information on management separately.

We sampled students and teachers to obtain *cross-sectional* information in grades 3 and 5 every year, as well as *longitudinal* information on the students who started grade 3 in 2013. Thus, in 2013, all students and teachers from grades 3 and 5 participated; in 2014, all students and teachers from grades 3 *to* 5 participated; and in 2015, all students and teachers from grades 3 and 5 participated. All principals in selected schools participated in the study.

3.3 Randomization

We randomly assigned the 105 sampled schools to one of three experimental groups, stratifying our randomization by school size to maximize statistical power. First, we grouped sampled schools into three strata by enrollment terciles. Then, we randomly assigned schools within each stratum to: (a) a "diagnostic feedback" or T1 group, in which we administered standardized tests in math and reading comprehension (in Spanish) at baseline and two follow-ups and made their results available to schools through user-friendly reports; (b) a "capacity building" or T2 group, in which we also provided schools with professional development workshops for school supervisors, principals, and teachers; or (c) a control group, in which we administered standardized tests only at the second follow-up. This process resulted in 30 T1 schools, 30 T2 schools, and 45 control schools.[10]

We randomly assigned schools to experimental groups in June of 2013 using administrative data for the 2013 school year provided by the ministry of education of the province. T1 and T2 schools were told that they would be part of the study in August of that year, but the list of control schools was not disclosed until late 2015, just before our last round of data collection, to minimize John Henry effects (i.e., control schools working harder than they otherwise would to compensate for not receiving any intervention). As we discuss in Section 3.4.3, we did not collect any data at these schools prior to this last round.

This setup allows us to estimate the effect of: (a) diagnostic feedback (comparing T1 to control schools in year 2 of the study); (b) combining diagnostic feedback with capacity building (comparing T2 to control schools in year 2); and (c) the value-added of capacity building, over and above diagnostic feedback (comparing T1 to T2 schools in years 1 and 2).

3.4 Treatment

Table 1 shows the timeline for the student assessments (on which our interventions were based) and the interventions. We discuss each experimental group in the sub-sections that follow.

[Insert Table 1 here.]

The school year in Argentina starts in February and ends in December. As the table shows, we administered the student assessments at the end of the 2013 and 2014 years and delivered the school reports based on those tests at the beginning of the 2014 and 2015 years. The professional development workshops took place during the 2014 and 2015 years.

3.4.1 Diagnostic feedback (T1) group

Schools assigned to the T1 group participated in the student assessments and received reports on the assessment results.[11] These reports were brief (about 10 pages) and had four

[10] Shortly after randomization, we had to drop one control school that the government had incorrectly categorized as public even though it was actually private.

[11] We discuss the grades that participated in each round of assessments in Section 4.

sections: (a) an introduction, which included a brief description the assessments and the share of participating students by grade and subject; (b) an overview of the school's average performance, which reported the school's average score by grade and subject,[12] the change in the school's average score from the previous year, the province's average score and change over the same period, and a comparison of the school's performance, vis-à-vis all other schools in the province and other urban and semi-urban schools; (c) an analysis of the distribution of the school's performance, which included a box-and-whiskers plot for the province for the last two years, an overlaid plot for the school, and another overlaid plot for each section in the school; and (d) a "traffic light" display of item-wise percent-correct results, organized by content and cognitive domains, displayed as red, yellow, or green based on proficiency cutoffs.[13]

As Table 1 indicates, some T1 schools participated in the workshops and visits designed for T2 schools. Thus, our impact estimates of T1 capture the effect of the T1 intervention as originally designed plus the (uneven participation in) school visits and a workshop on teaching geometry for teachers, or diagnostic feedback with minimal capacity building.

3.4.2 Capacity-building (T2) group

Schools assigned to the capacity-building (T2) group were assessed in 2013, 2014, and 2015. They were offered the reports described above; five workshops for supervisors, principals, and teachers; and two school visits. There were two workshops that explained the assessment results after each round of delivery of reports, one workshop on school improvement plans, one on quality assurance mechanisms, and one on geometry instruction. The first four workshops were offered to supervisors and principals and the last one to teachers. The school visits included a meeting with the principal and his/her leadership team, a classroom observation, and a meeting with the teaching staff. The workshops and visits were conducted by the ministry of education of the province, in collaboration with a local think tank. After each visit, the ministry prepared and shared a feedback report with the school, which included a diagnosis and some recommendations for improvement.[14]

As Table 1 shows, participation of T2 schools in some workshops and school visits was lower than expected. As we discuss in Section 5.2, we can exploit this variation in take-up to estimate the effect of receiving the components of T2 among a subset of schools.

[12] This score was scaled using a two-parameter Item Response Theory (IRT) model.

[13] A template of the report in English can be accessed at `http://bit.ly/2xrRaoc`.

[14] The design and content of the activities included in the workshops and school visits followed many of the recommendations in Boudett et al. (2005).

3.4.3 Control group

Schools assigned to the control group were only assessed in 2015. Schools in La Rioja have never participated in sub-national assessments, so administering the tests in 2013 and 2014 could have prompted behavioral responses from principals, teachers, and/or students due to increased monitoring that we wanted to avoid. We wanted to estimate the effect of administering the tests *and* either providing diagnostic feedback alone or with capacity building, as compared to "business-as-usual". None of the control schools received the school reports or workshops mentioned above.

3.4.4 Theory of change

Table 2 presents the theory of change of the interventions. Both T1 and T2 seek to address a key binding constraint: the lack of reliable, timely, and relevant student achievement data for management and instructional decisions. Additionally, T2 tries to address two potential obstacles that may mediate the ability of supervisors, principals, and teachers to use of student achievement data: the lack of capacity to analyze and/or act on these data.

[Insert Table 2 here.]

T1 and T2 aim to tackle the need for student achievement data through the administration of standardized tests and the dissemination of their results through annual reports. Additionally, T2 hopes to mitigate the potential lack of capacity of supervisors and principals to analyze the data through workshops that explain the results of assessments. T2 also aspires to address the potential lack of capacity of supervisors, principals, and teachers to act on these data through thematic workshops (e.g., on quality assurance and problem subjects/areas).

Both T1 and T2 seek to impact students' performance in school by: (a) lowering student absenteeism and tardiness (through increased student engagement and preparation); (b) improving student behavior (through more relevant and better instructional strategies); (c) lowering student repetition and dropout; and ultimately (d) increasing student achievement.

4 Data

As Table 3 shows, throughout the study, we administered student assessments of math and reading comprehension and surveys of students, teachers, and principals. We collected these data only in T1 and T2 schools in the first two years and in all sampled schools in the third year. We also collected administrative data from all schools prior to randomization and intervention monitoring data from all schools at the end of the study. Appendix A describes all instruments in detail. Below, we summarize the most important aspects of each instrument.

[Insert Table 3 here.]

4.1 Student assessments

We administered standardized tests of math and reading in all three years of the study. They were designed to assess what students ought to know and be able to do according to national and sub-national standards. They included 30 to 35 multiple-choice questions spanning a wide range of difficulty levels. The math test covered number properties, geometry, measurement, and probability and statistics and the reading test informative, narrative, and short texts.

We scaled the results to account for differences between items (specifically, their difficulty, capacity to distinguish between students of similar knowledge and skills, and propensity to be answered correctly by guessing) using a three-parameter Item Response Theory (IRT) model. We also included a common set of items across assessments of each grade and subject and used the IRT model to express the results from all three years on the same scale.

4.2 Student surveys

We also administered surveys of students in all three years of the study. In the first year, the surveys enquired about students' demographic characteristics, home environment, schooling trajectory, and study habits. In the second and third years, they also included questions on how frequently teachers engaged in certain activities (e.g., the teacher assigned homework or used the textbook) and students had positive interactions with teachers (e.g., the teacher gave students time to explain their ideas or checked that students understood the material).

4.3 Teacher surveys

We administered surveys of teachers in all three years of the study. In the first year, the surveys asked teachers about their demographic characteristics, education and experience, professional development, and teaching practices. In the second and third years, they also included questions on teachers' instructional practices, monitoring and evaluation practices at their schools, their job satisfaction, and the most important challenges they faced on the job.

4.4 Principal surveys

We also administered surveys of principals in the second and third years of the study. In both years, the surveys asked principals about their demographic characteristics, education and experience, professional development, and teaching practices, management practices, facilities and resources at their schools, and the most important challenges they faced on the job.

5 Empirical strategy

5.1 Intent-to-treat (ITT) effect

5.1.1 First-year effects

We estimate the effect of the offer (i.e., the intent-to-treat or ITT effect) of capacity building, over and above that of diagnostic feedback, after one year, by fitting:

$$Y_{ijk}^{t=1} = \alpha_l + \beta T2_k + \theta Y_{jk}^{t=0} + \varepsilon_{ijkl}^{t} \tag{1}$$

where $Y_{ijk}^{t=1}$ is the test score for student i in grade j in school k at the first follow-up, $Y_{jk}^{t=0}$ is the school-by-grade level average of that score at baseline,[15] α_l are school size (i.e., randomization strata) fixed effects, $T2$ is an indicator variable for schools assigned to T2, and ε_{ijkl} is the error term. The coefficient of interest is β, which indicates the magnitude of the value-added of capacity building after one year, with respect to diagnostic feedback. We adjust the standard errors to account for within-school correlations across students in outcomes. We test the sensitivity of our estimates to the inclusion of $Y_{jk}^{t=0}$.

5.1.2 Second-year effects

On the first year of the study, we estimate the ITT effect of diagnostic feedback, alone and combined with capacity building, by fitting the following model:

$$Y_{ijk}^{t=2} = \alpha_l + \beta_1 T1_k + \beta_2 T2_k + \lambda I_k^{t=0} + \varepsilon_{ijkl}^{t} \tag{2}$$

where $Y_{ijk}^{t=2}$ is the test score for student i in grade j in school k at the second follow-up, $T1$ is an indicator variable for schools assigned to T2, I_k is an index of school-level covariates from

[15] Unfortunately, students were not assigned unique IDs that allowed us to match their test scores to contemporaneous surveys, or to track their performance over time, so we cannot account for each student-level covariates or performance at baseline as we had originally planned.

administrative data at baseline,[16] and everything else is defined as above. The coefficients β_1 and β_2 indicate the magnitude of the effect of the offer of T1 and T2 after two years, with respect to business-as-usual operation. As above, we adjust the standard errors to account for clustering of outcomes within schools and across students. We test the sensitivity of our estimates to the inclusion of $I_k^{t=0}$. We also conduct two F-test to test the joint significance of β_1 and β_2, and to test that both coefficients are equal.

In equation (2), we cannot account for test scores at baseline because control schools were only assessed at the second follow-up (see Section 3.4.3). Yet, we can estimate the ITT effect of the offer of capacity building, over and above diagnostic feedback, after two years, by fitting:

$$Y_{ijk}^{t=2} = \alpha_l + \beta T2_k + \theta Y_{jk}^{t=0} + \varepsilon_{ijkl}^t \tag{3}$$

where everything is defined above and β indicates the magnitude of the value-added of capacity building after two years, with respect to diagnostic feedback. As above, we adjust the standard errors to account for clustering of outcomes within schools and across students and test the sensitivity of our estimates to the inclusion of $Y_{jk}^{t=0}$.

5.2 Local average treatment effect (LATE)

As Table 1 indicates, all T1 and T2 schools received school reports on both years of the study, but there was ample variation in the number of workshops and visits that schools received. This variation stemmed from non-compliance by T2 schools (e.g., absenteeism to workshops) and from cross-over by T1 schools (e.g., attendance to workshops for T2 schools).

Therefore, we also estimate the effect of receiving capacity building (i.e., the local average treatment effect or LATE) on "compliers" (i.e., schools that take up capacity-building activities when randomly assigned to them, but not otherwise), over and above diagnostic feedback, after two years, by fitting the following two-stage least-squares instrumental variables model:

$$\begin{aligned} A_k^{t=2} &= \eta_l + \gamma_1 T1_k + \gamma_2 T2_k + \gamma_3 I_k^{t=0} + \nu_{ijkl}^t \\ Y_{ijk}^{t=2} &= \sigma_l + \beta_1 T1_k + \beta_2 \hat{A}_k^{t=2} + \beta_3 I_k^{t=0} + \epsilon_{ijkl}^t \end{aligned} \tag{4}$$

where A_k is a measure of take-up of capacity building (the number of workshops and visits that the school received),[17] η_l and σ_l are randomization fixed effects, ν_{ijkl}^t and ϵ_{ijkl}^t are the

[16] This index is the first principal component from a principal component analysis of school variables (total enrollment, geographic location, management structure, and overage rate) collected at baseline. We cannot account for school-by-grade level average of test scores at baseline like we do in equation (1) because we did not administer tests at baseline in control schools.

[17] The first four workshops targeted principals and the fifth one targeted teachers. Therefore, we count a school as having received workshops #1-#4 if its principal attended, and we count that school as having received workshop #5 if at least one of its teachers attended.

error terms of the first and second stages of the model, and everything else is defined as above. The take up of capacity-building activities is instrumented by the random assignment of T2, hence ensuring that $\hat{A}_k^{t=2}$ is exogenous. The coefficient β_2 captures the marginal effect of participating in each capacity-building activity among compliers. As above, we adjust the standard errors to account for clustering of outcomes within schools and across students. We account for the two-step procedure estimating analytical standard errors (Wooldridge 2002).[18]

6 Results

6.1 Balancing checks

We can use the administrative data from 2013 to check that all three experimental groups were comparable at baseline. As Table 4 shows, there are no statistically significant differences between control and T2 schools, or between T1 and T2 schools. We only find a marginally statistically significant difference between control and T1 schools, indicating the former have a higher "overage rate" than the latter (i.e., a greater share of students who are one or more years above the theoretical age for their grade, either because they repeated a grade or dropped out of and then resumed schooling). Yet, given the number of tests that we run in this table, we would expect this difference to emerge simply by chance.

[Insert Table 4 here.]

To check whether the differences that we observe are indicative of differences in the equivalence of expectations across groups, we run a regression of the treatment dummy on all variables in Table 4 and the randomization fixed effects and test the joint significance of all coefficients using an F-test.[19] We cannot reject the null that there is no difference between any groups.

However, the administrative data to which we had access were reported at the school-level and they are weakly correlated with student achievement.[20] Therefore, it seems more appropriate

[18] We also estimate the dose-response relationship between the number of workshops or visits received and the outcomes of interest. This, however, is not equivalent to the LATE of each component of capacity building. A school may participate in workshops, visits, or both. Thus, when we use random assignment to instrument for take-up of one component (e.g., workshops), we leave the other one out (e.g., visits) and our estimation fails to meet the exclusion restriction (Angrist et al. 1996). These should hence be interpreted as associations.

[19] Each regression included only the two experimental groups of interest. For example, the regression in column (5) included only control and T1 schools.

[20] The first principal component from a principal component analysis of administrative variables (total enrollment, geographic location, management structure, and overage rate) had correlations of .04 and .08 with school-level average scores on the third grade math and reading tests and correlations of .05 and .14 with the fifth grade math and reading tests.

to use the data collected at baseline to compare T1 and T2 schools.[21] A student-level t-test comparing the IRT-scaled scores at baseline across groups indicates that students at T1 schools already performed better in all grades and subjects at baseline, and those differences are statistically significant at least at the 5% level in all cases (Figure B.1). These differences are larger among students in grade 3 (.18σ in math and .21σ in reading) than in grade 5 (.09σ in math and .15σ in reading).[22] Yet, they are not statistically significant if we use a regression that accounts for the randomization fixed effects and the clustering of test scores (Table B.2). If we run two regressions of the treatment dummy on test scores and randomization fixed effects (one regression per grade), and test the joint significance of all coefficients using an F-test, we cannot reject the null that there are no differences between these groups.

We can also use the student and teacher surveys from baseline to compare T1 and T2 schools. If we run a regressions of the treatment dummy on either the student or teacher variables and the randomization fixed effects, and test the joint significance of all coefficients using an F-test, we can reject the null that there are no differences between these groups. In both cases, the difference is statistically significant at least at the 5% level (Tables B.3-B.4).

All of this evidence suggests that, by chance, T2 schools, were at a disadvantage with respect to T1 schools at baseline. As we discuss below, this disadvantage seems to explain the differences in the impact estimates of these interventions.

6.2 Treatment dosage

We can use the data from the principal surveys and intervention monitoring from 2015 to compare treatment dosage at schools in all three experimental groups by the end of the study. As expected, principals at T1 and T2 schools are far more likely to report engaging in the behaviors targeted by the interventions, such as administering standardized tests, tracking their results over time, and comparing their results with those of the province or other schools (Table B.5). Also as expected, according to the intervention monitoring data, all control schools received one report (after endline), whereas all treatment schools received three reports (two before and one after endline). T1 schools received more workshops and visits than control schools (even if neither T1 nor control schools were supposed to receive any) and T2 schools received more workshops and visits than control and T1 schools. Specifically, out of a total of 5 workshops and 2 school visits, the average T1 school received .23 workshops and .7 school visits while T2 school received 3.4 workshops and 1.5 visits over the two years of the study.

[21] As discussed in Section 3.4.3, we did not collect data from control schools at baseline. We collected baseline data after randomization, but we had not notified schools of their assignment, so we have no reasons to believe that the "stable unit treatment value assumption" (Imbens and Wooldridge 2009) was violated.

[22] These differences are driven by six schools in grade 3 and eight schools in grade 5 where one student performs considerably below their peers.

6.3 Average ITT effects

6.3.1 First-year effects

Table 5 presents the ITT effects of capacity building on test scores, over and above that of diagnostic feedback, after one year. As the table indicates, we cannot reject the null that T1 and T2 schools performed at the same level in 2014. In fact, the coefficient on the T2 dummy is negative, but small-to-moderate (between -.14 and -.02σ) and never statistically significant.

[Insert Table 5 here.]

Two sets of results in this table suggest that the differences in student achievement between T1 and T2 schools at baseline might explain why the coefficients on the T2 dummy are negative.[23] First, the magnitude of the negative coefficients is larger in grade 3, where the disadvantage of T2 schools was larger, than in grade 5. Second, once we account for school-by-grade level averages of test scores at baseline, the coefficients become less negative. These results indicate that capacity building did not add value to diagnostic feedback after a year.[24]

6.3.2 Second-year effects

Table 6 presents the ITT effects of diagnostic feedback on test scores, alone and combined with capacity building, after two years. As the table indicates, T1 schools outperformed control schools by .34 and .36σ in grade 3 math and reading, and by .28 and .38σ in grade 5 math and reading. All effects are statistically significant at the 1% level, except for the one on grade 5 math, which is statistically significant at the 5% level. Their sign and magnitude are robust to the inclusion of school-level covariates from administrative data at baseline.

[Insert Table 6 here.]

As the table also shows, T2 schools performed on par with control schools in grade 3 and they outperformed controls in grade 5 math and reading by .21 and .19σ. This is once again consistent with the differences in student achievement between T1 and T2 schools at baseline: we are more likely to observe larger and statistically significant effects of T2 in the grade at which the disadvantage of T2 schools was smaller. In fact, as column 7 indicates, except in grade 3 reading, we cannot reject the null that the coefficients on T1 and T2 are equal.

[23] This result would otherwise be surprising because T2 schools received capacity building *in addition to* the diagnostic feedback that T1 schools received.

[24] As Figure B.1 shows, the mean performance of T1 and T2 schools is very similar by 2014, even if T2 schools had performed at a much lower level in 2013.

The inclusion of an index of school-level covariates from administrative data at baseline does little to account for imbalances at baseline between T1 and T2 schools because, as we mentioned in Section 6.1, they are poorly correlated with achievement. If we limit our analysis to T1 and T2 schools and account for school-by-grade level averages of test scores at baseline using equation (3), the coefficients on T2 again become less negative (Table B.6).

The impact of diagnostic feedback can be observed across content and cognitive domains. Students at T1 schools outperformed their counterparts at control schools in nearly all content domains in math (i.e., numbers, geometry, measurement, and statistics) and reading (i.e., informative texts, narrative texts, and short texts) (Tables B.7-B.8), and in almost all cognitive domains in math (i.e., communicating, knowing, solving algorithms, and solving problems) and reading (i.e., extracting explicit and implicit information, analyzing texts, and reflecting) (Tables B.9-B.10).[25] We do not find this generalized impact among T2 schools, but in nearly all cases, we cannot reject the null that the coefficients on T1 and T2 schools are equal.

6.4 Potential mechanisms

6.4.1 School management

One way in which diagnostic feedback and capacity building may impact student achievement is by encouraging principals to use assessment results to inform school management decisions (Rockoff et al. 2012). We use the survey of principals from 2015 to check whether those in T1 and T2 schools were more likely to use results for this purpose than their control counterparts.

We find clear evidence that diagnostic feedback, alone and combined with capacity building, influenced school management practices. Table 7 presents the ITT effects of both interventions after two years. As the table indicates, principals at T1 and T2 schools were far more likely than their control peers to report using test score results to inform management decisions (e.g., setting goals for the school, making changes to the curriculum, or evaluating teachers). Interestingly, however, principals did not report using test results for student ability tracking. They were also more likely to report making the results available to parents and the public, even if this was neither required nor encouraged by the interventions.

[Insert Table 7 here.]

[25] The estimates that account for school-level covariates at baseline are nearly identical and are available from the authors upon request.

6.4.2 Classroom instruction

Another way in which diagnostic feedback and capacity building may impact achievement is by encouraging teachers to change their instructional practices. For example, they may engage in more activities during their lessons and/or improve their interactions with students.

We administered student surveys instead of classroom observations to measure the impact of the interventions on instruction for two main reasons. First, the type of observations that can be administered by enumerators at scale in developing countries are susceptible to Hawthorne effects (i.e., teachers exerting more effort on the day of the observation than on other days), even when they are unannounced (see, for example, Muralidharan and Sundararaman 2010). Second, whereas observations typically rely on a single occasion and rater, student surveys can draw on 20 to 30 raters (i.e., students) and include questions about different time periods (e.g., the weeks preceding the survey or the entire school year) (Kane and Staiger 2011, 2012).

We included two sets of questions on instruction in our second follow-up student survey.[26] We asked students to indicate how frequently their math and Spanish teachers engaged in a number of activities in the two weeks prior to the assessments, using a Likert-type scale that ranged from 1 ("never or only once") to 5 ("5 times or more"). We also asked them to indicate how frequently they had a series of positive interactions with the teacher and/or the material throughout the school year, using a similar scale that ranged from 1 ("never") to 5 ("always"). Thus, we measured both teacher activity and the quality of student-teacher interactions.

We find clear evidence that diagnostic feedback increased teacher activity. Table 8 presents the ITT effects of diagnostic feedback and capacity building on teacher activity after two years. As the table indicates, students at T1 schools were more likely than their control counterparts to report that their math and Spanish teachers assigned and graded homework, copied from the blackboard, and explained a topic. They were also more likely to report that they solved problems during math class and used a textbook or do a dictation during Spanish lessons. This is not the case in T2 schools. In fact, students at these schools were less likely than their control peers to report that their math teachers asked them to work in groups or graded their homework, and that their Spanish teacher graded their homework.[27]

[Insert Table 8 here.]

We also find clear evidence that diagnostic feedback improved student-teacher interactions. As Table 9 indicates, students at T1 schools reported to have positive interactions with their

[26] For details and a link to the instrument, see Appendix A.

[27] All of these effects remain statistically significant if we account for school-level covariates at baseline. Results available from the authors upon request.

teachers and the material on nearly all indicators. This suggests that teachers at T1 schools were not simply engaging in more activities; they also improved the quality of their instruction. This was not the case in T2 schools, which resembled control schools on all indicators.

[Insert Table 9 here.]

As our review of prior research in Section 2 had anticipated, diagnostic feedback mostly impacted instruction through the intensive rather than the extensive margin of teacher effort. We asked students how frequently their teacher had attended school, arrived on time, and started or ended lessons on time during the school year, on a 1 ("never") to 5 ("always") scale. Teachers in T1 and T2 schools were no more likely to go to work or arrive on time. (The sign of the coefficient on T1 schools is consistently negative, but statistically insignificant). However, they were less likely to end class or leave school early (Table B.11).[28] Therefore, as we had expected, teachers exerted more effort conditional on attending to school.

6.5 Heterogeneous ITT effects

6.5.1 School size

The effects of diagnostic feedback and capacity building may vary by school size. Specifically, prior evidence suggests smaller schools should fare better (Imberman and Lovenheim 2015). This may occur for two reasons. First, the effort of each individual teacher matters more in a small than in a large school. Therefore, it is more costly (e.g., in terms of self-perception and reputation) for a teacher to "free-ride" (i.e., exert low effort and rely on that of that of his/her peers to improve student achievement). Second, each teacher is also easier to monitor in a small than in a large school because there are fewer teachers to monitor. Thus, it is more difficult for teachers to "shirk" (i.e., exert low effort and expect no repercussions).

We test for heterogeneous effects in three ways. First, we use a version of equation (2) that includes the total enrollment at each school and its interaction with the treatment dummies. Second, we do the same using enrollment at each grade instead of total enrollment. Third, we use a version of equation (2) that includes the enrollment tercile (i.e., randomization strata) fixed effects for small and medium schools and their interactions with the treatment dummies.

We find little evidence of heterogeneity by school size. In the first and second specifications, the coefficients on the interaction terms are consistently estimated to be around zero and statistically insignificant (Tables B.12-B.13). In the third specification, results differ by treatment. The coefficients on the interactions with the T1 dummy are consistently negative,

[28] They were also less likely to start class early, but this would occur if they are not ending classes early.

but only statistically significant in the case of grade 3 reading (which is to be expected, given the number of tests that we are running).[29] The coefficients on the interactions with the T2 dummy vary widely in sign and magnitude across subjects and grades (Table B.14).

6.5.2 School location

Effects may also vary by schools' geographic location. Yet, it is not clear whether urban or semi-urban schools should benefit more from the interventions. Urban schools are exposed to more external pressures (e.g., demands from more parents, competition with other schools, proximity to government officials), which could make them more responsive to information and/or capacity, but semi-urban schools are smaller and thus potentially easier to change.

We test for heterogeneous effects by fitting a version of equation (2) with a dummy for urban schools and the interactions between that dummy and each of the treatment dummies. The coefficients on the interactions are positive, but mostly statistically insignificant (Table B.15).

6.6 Threats to validity

6.6.1 Test familiarity

It is possible that the effect of diagnostic feedback on test scores was driven by test familiarity. By 2015, T1 and T2 schools had administered student assessments for three years in a row, whereas control schools administered these assessments for the first time on that year. Thus, T1 schools may have outperformed control schools in 2015 because their teachers and students were more familiar with the assessments, not because their students had learned more.

This explanation is unlikely to drive our effects. First, if we believe that it is *students'* familiarity with the tests that matters, it is hard to explain why T1 schools outperformed their control counterparts both in grade 3 (where students were taking the tests for the first time) and grade 5 (where students had taken the tests twice before) (see Table 6 above). Second, to examine whether *teachers'* familiarity with the tests matters (e.g., because they adjusted their instruction based on the items they saw in 2013 and 2014),[30] we run our impact estimations separately for "familiar" items (i.e., included on previous rounds of assessments) and "unfamiliar" items (i.e., not included on previous assessments). We find that T1 schools outperformed control schools on both types of items (Table B.16).

[29] These negative coefficients suggest that, in large schools, it may be too difficult for principals to monitor student achievement directly, and reports are a useful input for school management.

[30] Teachers were not allowed to monitor the assessments in the section that they taught. However, primary school teachers in La Rioja are typically "homeroom" teachers (i.e., they teach all subjects to their students). Therefore, a math teacher could have seen the items on the math test when he/she was monitoring the assessment in a different section. Teachers were not given or allowed to keep copies of the assessments.

6.6.2 Student absenteeism

It is also possible that the effect of diagnostic feedback on test scores was due to non-random student absenteeism. Specifically, T1 schools may have discouraged low-performing students from going to school on the day of the assessments. We cannot determine whether this occurred because we do not have a panel of test scores. Yet, we collected data on student absenteeism from each school's register in 2015 to test whether the effects on T1 schools are driven by the *share* of students who were absent on testing day. We find that our impact estimates remain virtually unchanged once we account for each school's share of absent students (Table B.17).

6.6.3 Contamination

Finally, it is also possible that supervisors who were responsible for control and treatment schools used the inputs and/or practices from the interventions at control schools. If so, our estimates would be a lower-bound of the true effect of diagnostic feedback. We collected the list of schools for which each supervisor was responsible to test whether effects are lower for supervisors that oversee control and T1 schools. We do not find this is the case (Table B.18).

6.7 LATE of capacity building

6.7.1 Second-year effects

Table 10 presents the LATE of participating in each activity (i.e., workshop or visit) of the capacity-building group, after two years. As the table indicates, being assigned to this group is associated with nearly five workshops or visits during the two years of the program, and each activity is positively associated with student achievement gains (by .02 to .05σ, depending on the subject and grade), but these associations are not statistically significant. These results are consistent with the dose-response relationships between the number of workshops and visits and student achievement, when estimated separately (Tables B.19-B.20).

[Insert Table 10 here.]

These results also contribute to our understanding of why T2 schools did not perform better. First, they were at a disadvantage with respect to T1 schools at baseline (see Section 6.1). Second, fewer than expected T2 schools participated in capacity-building activities and participation in each activity did little to improve student achievement.

7 Discussion

This paper presented experimental evidence on the impact of providing diagnostic feedback, alone and combined with capacity building, to public primary schools based on a sub-national large-scale student assessment in the Province of La Rioja, Argentina. We found moderate to large positive effects of diagnostic feedback in third and fifth grades, in math and reading. We found smaller effects of capacity building in fifth grade, but we could not discard the possibility that diagnostic feedback alone had the same effect when combined with capacity building. The interventions were equally effective in small and large schools, and in urban and semi-urban schools. Our impact estimates were not driven by student absenteeism, familiarity with test items, or contamination through school supervisors.

The impact of diagnostic feedback demonstrates the potential of large-scale assessments to inform school management and classroom instruction. Upon receiving the assessment results, principals used them as an input for school management decisions and teachers resorted to more pedagogical strategies and improved their interactions with students. Our results seem to confirm our interpretation of prior studies, which suggest that diagnostic feedback may be less useful in lower-middle-income countries like India, where the binding constraint of school systems is the *extensive* margin of worker effort (i.e., getting teachers to go to school and teach for the full lesson), and more useful in upper-middle-income countries like Mexico or Argentina, where the binding constraint is the *intensive* margin of worker effort (i.e., getting teachers to increase their effort when they go to school) (de Hoyos et al. 2017; Muralidharan and Sundararaman 2010).

The uneven impact of capacity building illustrates the challenges of implementing meaningful professional development in developing countries. Schools assigned to receive capacity building participated in fewer workshops and visits than originally planned. Additionally, each activity (i.e., workshop or visit) had a positive but limited and statistically insignificant impact on achievement. Our results are consistent with those of evaluations of professional development programs in developing countries, which have also found low take-up and limited effects on learning (see, for example, Angrist and Lavy 2001; Yoshikawa et al. 2015; Zhang et al. 2013).

Overall, our results suggest that diagnostic feedback may be sufficient to elicit improvements in the management and instruction of public schools.[31] While complementing such feedback with capacity building may seem intuitive, it is challenging to implement faithfully in practice. Yet, whether diagnostic feedback has a similar effect in other school systems will depend on the extent to which those systems share the distinctive features of La Rioja, including the limited access that public schools had to student achievement data prior to the experiment

[31] As stated above, some schools assigned to diagnostic feedback participated in some workshops and school visits, so this should be interpreted as the impact of diagnostic feedback with *minimal* capacity building.

(which means the school reports add more value than they would in systems where student achievement data is more readily available) and the relatively small and manageable size of the system (which makes it easy to implement a quality assurance mechanism). It will also depend on the extent to which the reforms adopted by those systems include some of the defining features of the feedback intervention in La Rioja, such as assessments that cover all students and are comparable over time (which allow for comparisons of schools' performance over time) and user-friendly and timely reports that present information not only on achievement levels and changes, but also on content and cognitive domains (which allow teachers to identify problem areas). These characteristics of the context and intervention are likely to be important mediators of the impact of similar reforms in other upper-middle-income countries.

References

Abadzi, H. (2007). Absenteeism and beyond: Instructional time loss and consequences. (Policy Research Working Paper No. 4376). The World Bank. Washington, DC.

Andrabi, T., J. Das, and A. I. Khwaja (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review 107*(6), 1535–1563.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Angrist, J. D. and V. Lavy (2001). Does teacher training affect pupil learning? evidence from matched comparisons in jerusalem public schools. *Journal of Labor Economics 19*(2).

Banerjee, A. V., R. Banerji, E. Duflo, and M. Walton (2011). Effective pedagogies and a resistant education system: Experimental evidence on interventions to improve basic skills in rural India. *Unpublished manuscript*. New Delhi, India: Abdul Latif Jameel Poverty Action Lab (J-PAL).

Boudett, K. P., E. A. City, and R. J. Murnane (2005). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning*. Cambridge, MA: Harvard Education Press.

Bruns, B. and J. Luque (2014). *Great teachers: How to raise student learning in Latin America and the Caribbean*. Washington, DC: The World Bank.

Busso, M., M. Bassi, and J. S. Muñoz (2013). Is the glass half empty or half full? School enrollment, graduation, and dropout rates in Latin America. (IDB Working Paper Series No. IDB-WP-462). Washington, DC: Inter-American Development Bank.

Camargo, B., R. Camelo, S. Firpo, and V. Ponczek (2011). Test score disclosure and school performance. (Sao Paulo School of Economics Working Paper No. 11/2011). Center for Applied Economics. Sao Paulo, Brazil.

Chaudhury, N., J. Hammer, M. Kremer, K. Muralidharan, and F. H. Rogers (2006). Missing in action: Teacher and health worker absence in developing countries. *The Journal of Economic Perspectives 20*(1), 91–116.

Cheng, X. and C. Gale (2014). National assessments mapping metadata. Washington, DC: FHI 360. Retrieved from: `http://bit.ly/2yxBeBd`.

de Hoyos, R., V. A. García-Moreno, and H. A. Patrinos (2017). The impact of an accountability intervention with diagnostic feedback: Evidence from Mexico. *Economics of Education Review 58*, 123–140.

de Hoyos, R., P. A. Holland, and S. Troiano (2015). Understanding the trends in learning outcomes in Argentina, 2000 to 2012. (Policy Research Working Paper No. 7518). The World Bank. Washington.

DiNIECE (2013). Redefiniciones normativas y desafíos de la educación secundaria en Argentina. Acuerdos federales en un sistema descentralizado. La educación en debate. Buenos Aires, Argentina: Dirección Nacional de Información y Evaluación de la Calidad Educativa (DiNIECE).

DiNIEE (2015). Anuario Estadístico 2015. Buenos Aires, Argentina: Dirección Nacional de Información de la Calidad Educativa (DiNIECE).

Duflo, E., J. Berry, S. Mukerji, and M. Shotland (2015). A wide angle view of learning: Evaluation of the CCE and LEP programmes in Haryana, India. (Impact Evaluation Report No. 22). International Initiative for Impact Evaluation (3ie). New Delhi, India.

Duro, E., M. Scasso, S. Bonelli, A. Hoszowski, R. Cortés, C. Giacometti, and V. Volman (2016). Operativo Nacional de Evaluación (ONE) 2013: Diagnóstico y consideraciones metodológicas necesarias para el análisis y difusión de sus resultados. Ciudad Autónoma de Buenos Aires, Argentina: Secretaría de Evaluación Educativa, Ministerio de Educación y Deportes de la Nación.

Ferrer, G. (2006). *Educational assessment systems in Latin America: Current practice and future challenges.* Partnership for Educational Revitalization in the Americas (PREAL).

Ferrer, G. and A. Fiszbein (2015). What has happened with learning assessment systems in Latin America? Lessons from the last decade of experience. Washington, DC: The World Bank.

Ganimian, A. J. (2013). No logramos mejorar: Informe sobre el desempeño de Argentina en el Programa para la Evaluación Internacional de Alumnos (PISA) 2012. Buenos Aires, Argentina: Proyecto Educar 2050.

Ganimian, A. J. (2014). Avances y desafíos pendientes: Informe sobre el desempeño de Argentina en el Tercer Estudio Regional Comparativo y Explicativo (TERCE) del 2013. Buenos Aires, Argentina: Proyecto Educar 2050.

Ganimian, A. J. (2015). El termómetro educativo: Informe sobre el desempeño de Argentina en los Operativos Nacionales de Evaluación (ONE) 2005-2013. Buenos Aires, Argentina: Proyecto Educar 2050.

Ganimian, A. J. and D. M. Koretz (2017). Dataset of international large-scale assessments. Last updated: February 8, 2017. Cambridge, MA: Harvard Graduate School of Education.

Hanushek, E. A. and L. Woessmann (2007). Education quality and economic growth. Washington, DC: The World Bank.

Hanushek, E. A. and L. Woessmann (2010). *The high cost of low educational performance: The long-run economic impact of improving PISA outcomes.* Paris, France: Organisation for Economic Co-operation and Development.

Harris, D. (2005). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice 8*(1), 35–41.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments of the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

Imberman, S. A. and M. F. Lovenheim (2015). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *The Review of Economics and Statistics 2*(97), 364–386.

Kane, T. J. and D. O. Staiger (2011). Learning about teaching: Initial findings from the Measures of Effective Teaching project. Measures of Effective Teaching Project. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J. and D. O. Staiger (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Measures of Effective Teaching Project. Seattle, WA: Bill and Melinda Gates Foundation.

Kremer, M., N. Chaudhury, F. H. Rogers, K. Muralidharan, and J. Hammer (2005). Teacher absence in india: A snapshot. *Journal of the European Economic Association 3*(2-3), 658–667.

Mizala, A. and M. Urquiola (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics 103*, 313–335.

Muralidharan, K., J. Das, A. Holla, and A. Mohpal (2017). The fiscal cost of weak governance: Evidence from teacher absence in India. *Journal of Public Economics*.

Muralidharan, K. and V. Sundararaman (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from India. *The Economic Journal 120*(F187-F203).

Piper, B. and M. Korda (2011). EGRA plus: Liberia. Program evaluation report. Unpublished manuscript. RTI International. Research Triangle Park, NC.

Pritchett, L. (2013). *The rebirth of education: Schooling ain't learning.* Washington, DC: Center for Global Development.

Rockoff, J. E., D. O. Staiger, T. J. Kane, and E. S. Taylor (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *The American Economic Review 102*(7), 3184–3213.

Sankar, D. and T. Linden (2014). How much and what kind of teaching is there in elementary education in India? Evidence from three states. (South Asia Human Development Sector Report No. 67). Washington, DC: The World Bank.

SEE-MEDN (2016). Aprender 2016. Ciudad Autónoma de Buenos Aires, Argentina: Secretaría de Evaluación Educativa, Ministerio de Educación y Deportes de la Nación.

SEE-MEDN (2017). Aprender 2016: Análisis de desempeños por capacidades y contenidos. Nivel primario. (Serie de documentos técnicos, Nro. 7.) Ciudad Autónoma de Buenos Aires: Secretaría de Evaluación Educativa. Ministerio de Educación y Deportes de la Nación.

Stallings, J. A., S. L. Knight, and D. Markham (2014). Using the Stallings observation system to investigate time on task in four countries. *Unpublished manuscript.* Washington, DC: The World Bank.

UNGA (2000). A/RES/55/2. Resolution adopted by the General Assembly on 18 September 2000. New York, NY: United Nations General Assembly.

UNGA (2015). A/RES/70/1. Resolution adopted by the General Assembly on 25 September 2015. New York, NY: United Nations General Assembly.

Wooldridge, J. M. (2002). *Econometric analysis of cross-section and panel data.* Cambridge, MA: MIT Press.

World Bank (2016). What is happening inside classrooms in Indian secondary schools? A time on task study in Madhya Pradesh and Tamil Nadu. *Unpublished manuscript.* New Delhi, India: The World Bank.

World Bank (2017). World Development Report 2017: Learning to realize education's promise. Washington, DC: The World Bank.

Yen, W. M. and A. R. Fitzpatrick (2006). Item response theory. In Brennan, R. (Ed.) *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Yoshikawa, H., D. Leyva, C. E. Snow, E. Treviño, M. C. Arbour, M. C. Barata, C. Weiland, C. Gómez, L. Moreno, A. Rolla, and N. D'Sa (2015). Experimental impacts of a teacher professional development program in chile on preschool classroom quality and child outcomes. *Journal of Developmental Psychology 51*, 309–322.

Zhang, L., F. Lai, X. Pang, H. Yi, and S. Rozelle (2013). The impact of teacher training on teacher and student outcomes: evidence from a randomised experiment in beijing migrant schools. *Journal of development effectiveness 5*(3), 339–358.

Table 1: Timeline of the intervention

			School participation rates		
Year	Month	Event	Control	T1: diagnostic feedback	T2: capacity building
2013	Feb	*School year starts*			
	Oct	Assessment of grades 3 and 5	-	100%	100%
2014	Feb	*School year starts*			
	Mar	Delivery of school reports	-	100%	100%
		Workshop # 1: Assessment results	-	-	53%
	Apr	School visit #1	-	40%	60%
	May	Workshop # 2: School improvement plans	-	-	90%
	Sep	Workshop # 3: Quality assurance	-	-	87%
	Nov	Assessment of grades 3, 4, and 5	-	100%	100%
2015	Feb	*School year starts*			
	Apr	Delivery of school reports	-	100%	100%
		Workshop # 4: Assessment results	-	-	97%
	Jun	School visit #2	-	33%	87%
	Sep	Workshop # 5: Teaching geometry	-	23%	20%
	Oct	Assessment of grades 3 and 5	100%	100%	100%

Notes: Participation rates indicate the percentage of schools in each experimental group that participated in each event.

Table 2: Theory of change

Need	Inputs	Outputs	Outcomes	Impact
• **Principals and teachers lack data to improve management and instruction** (to better manage school resources and tailor instruction)	• Administration of standardized tests (T1 and T2) • Annual reports are distributed (T1 and T2)	• Principals improve management practices and teachers improve instructional practices	• Principals recruit better-qualified teachers • Principals shift time from administrative to instructional tasks • Teachers spend more time in the classroom • Teachers experiment with classroom strategies • Students rate their teachers more favorably	• Lower student absenteeism and tardiness • Better student behavior • Lower student repetition and dropout • Higher student achievement
• **Supervisors and principals need assistance analyzing student achievement data** (to make sense of data and identify areas for improvement)	• Workshops on assessment results and use of results (T2)	• Supervisors and principals analyze student achievement data	• School's performance is tracked over time and compared across schools, province, and country • Principals share data with parents and public	
• **Supervisors, principals, and teachers lack capacity to act on student achievement data** (to devise school improvement strategies)	• Workshops on school improvement, quality assurance, and problem subjects/areas (T2) • Visits to schools (T2)	• Supervisors, principals, and teachers make more decisions based on test results	• Principals develop school improvement plans • Principals and teachers are evaluated partly based on test results	

Table 3: Timeline of data collection

			School participation rates		
Year	Month	Event	Control	T1: diagnostic feedback	T2: capacity building
2013	Feb	*School year starts*			
	Mar	Administrative data on schools	100%	100%	100%
	Oct	Assessments of grade 3 and 5	-	100%	100%
		Survey of teachers in grades 3 and 5	-	100%	100%
		Survey of students in grades 3 and 5	-	100%	100%
2014	Feb	*School year starts*			
	Nov	Assessment of grades 3-5	-	100%	100%
		Survey of teachers in grades 3-5	-	100%	93%
		Survey of students in grades 3-5	-	100%	100%
		Survey of principals	-	100%	93%
2015	Feb	*School year starts*			
	Oct	Assessment of grades 3 and 5	100%	100%	100%
		Survey of teachers in grades 3 and 5	100%	100%	100%
		Survey of students in grades 3 and 5	100%	100%	100%
		Survey of principals	100%	100%	100%

Notes: Participation rates indicate the percentage of schools in each experimental group that participated in each round of data collection.

Table 4: Balancing checks on school administrative data (2013)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variable	All	C	T1	T2	T1-C	T2-C	T2-T1	N
Enrollment - Total	332.615	322.311	356.667	323.724	2.174	-.476	-4.597	104
	(265.535)	(323.729)	(207.248)	(222.099)	(40.681)	(42.145)	(27.697)	
Enrollment - Grade 3	48.308	45.756	52.633	47.793	2.074	1.768	-.405	104
	(39.663)	(45.094)	(32.501)	(38.295)	(5.692)	(6.707)	(5.36)	
Enrollment - Grade 5	47.481	46.6	49.733	46.517	-1.345	-.335	.675	104
	(39.83)	(50.187)	(29.175)	(31.44)	(6.878)	(6.76)	(4.28)	
Urban school	.606	.578	.667	.586	.042	.012	-.027	104
	(.491)	(.499)	(.479)	(.501)	(.088)	(.1)	(.101)	
Semi-urban school	.39	.422	.333	.4	-.042	-.012	.027	105
	(.49)	(.499)	(.479)	(.498)	(.088)	(.1)	(.101)	
School has principal and vice-principal	.781	.778	.867	.7	.053	-.05	-.108	105
	(.416)	(.42)	(.346)	(.466)	(.076)	(.086)	(.093)	
School only has principal	.105	.111	.1	.1	.007	-.009	-.017	105
	(.308)	(.318)	(.305)	(.305)	(.07)	(.068)	(.074)	
Overage - Grade 3	7.442	6.638	6.979	9.056	.356	2.619	2.327	103
	(8.611)	(7.836)	(7.726)	(10.409)	(1.803)	(2.287)	(2.291)	
Overage - Grade 5	11.436	13.588	8.958	10.828	-4.438*	-2.387	2.514	103
	(12.326)	(14.776)	(8.486)	(11.515)	(2.628)	(3.136)	(2.56)	
F-statistic					1.569	.988	1.126	
p-value					.133	.462	.359	

Notes: (1) The table shows the mean and standard deviations of all schools in the sample (column 1), C schools (column 2), T1 schools (column 3), and T2 schools (column 4). It also tests for differences across C and T1 schools (column 5), C and T2 schools (column 6), T1 and T2 schools (column 7), and shows the number of non-missing observations (column 8). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 5-7 account for clustering at the school level. (4) All specifications include randomization strata fixed effects.

Table 5: ITT effects on student achievement data (2014)

	(1)	(2)	(3)	(4)
	Constant	T2	N	Controls?
Panel A. Grade 3				
Math (scaled)	.379**	-.092	2618	N
	(.155)	(.136)		
	.353**	-.006	2618	Y
	(.15)	(.094)		
Reading (scaled)	.157	-.146	2514	N
	(.19)	(.114)		
	.118	-.053	2514	Y
	(.223)	(.079)		
Panel B. Grade 5				
Math (scaled)	.092	-.023	2562	N
	(.176)	(.14)		
	-.005	-.002	2562	Y
	(.179)	(.116)		
Reading (scaled)	-.028	-.027	2465	N
	(.079)	(.113)		
	.019	.042	2465	Y
	(.12)	(.078)		

Notes: (1) The table displays a different regression in every line, using two versions of equation (1): one that does not include school-by-grade level averages of test scores at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and T2 dummy (column 2), the number of non-missing observations (column 3), and whether the school-by-grade level average of test scores at baseline was included (column 4). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 account for clustering at the school level. (4) All estimations include randomization strata fixed effects.

Table 6: ITT effects on student achievement data (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
						F-tests	
	Constant	T1	T2	N	Controls?	$\beta_1 = \beta_2 = 0$	$\beta_1 = \beta_2$
Panel A. Grade 3							
Math (scaled)	.03	.343***	.156	3882	N	4.194	1.268
	(.168)	(.119)	(.154)			(.018)	(.263)
	.024	.34***	.154	3882	Y	3.864	1.27
	(.17)	(.122)	(.159)			(.024)	(.262)
Reading (scaled)	-.074	.356***	.108	3993	N	5.401	2.97
	(.141)	(.109)	(.131)			(.006)	(.088)
	-.045	.372***	.122	3993	Y	5.745	3.029
	(.144)	(.11)	(.134)			(.004)	(.085)
Panel B. Grade 5							
Math (scaled)	.191	.284**	.214*	4150	N	3.578	.23
	(.139)	(.116)	(.124)			(.032)	(.633)
	.189	.283**	.214*	4150	Y	3.21	.23
	(.137)	(.12)	(.128)			(.045)	(.633)
Reading (scaled)	-.056	.378***	.188*	4260	N	5.855	1.865
	(.105)	(.114)	(.112)			(.004)	(.175)
	-.059	.377***	.187	4260	Y	5.306	1.865
	(.102)	(.118)	(.117)			(.006)	(.175)

Notes: (1) The table displays a different regression in every line, using two versions of equation (2): one that does not include an index of school-level covariates from administrative data at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), the number of non-missing observations (column 4), and whether the index of school-level covariates at baseline was included (column 5). It also shows the results from two F-tests: one testing whether the coefficients on the T1 and T2 dummies were jointly statistically significant (column 6) and another one testing whether the coefficient on the T1 dummy is statistically significantly different from the coefficient on the T2 dummy (column 7). In both columns, the F-statistic is shown with its associated p-value (in parentheses). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.

Table 7: ITT effects on principal-reported management practices (2015)

	(1) Constant	(2) T1	(3) T2	(4) N
School set goals based on tests	.424***	.458***	.467***	84
	(.109)	(.1)	(.102)	
Curriculum changed based on tests	.659***	.286***	.221**	85
	(.105)	(.083)	(.098)	
Principal evaluated based on tests	.41***	.276**	.375***	83
	(.112)	(.133)	(.126)	
Teacher evaluated based on tests	.469***	.222*	.21	86
	(.119)	(.127)	(.133)	
Students tracked based on tests	.09	.084	.009	82
	(.062)	(.1)	(.085)	
Parents were informed of test results	.453***	.396***	.469***	88
	(.105)	(.109)	(.104)	
Test results were made public	.279**	.288**	.303**	83
	(.109)	(.127)	(.129)	

Notes: (1) The table displays a different regression in every line, using two versions of equation (1): one that does not include school-by-grade level averages of test scores at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), and the number of non-missing observations (column 4). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 to 3 account for clustering at the school level. (4) All estimations include randomization strata fixed effects.

Table 8: ITT effects on student-reported teacher activity (2015)

	(1) Constant	(2) T1	(3) T2	(4) N
Panel A. Math				
I used a textbook	2.76***	.115	.068	7039
	(.133)	(.092)	(.115)	
My teacher assigned me homework	3.714***	.128*	-.005	7318
	(.111)	(.076)	(.101)	
I copied from the blackboard	3.606***	.161**	.008	7140
	(.103)	(.069)	(.071)	
I solved problems	3.838***	.237***	.055	7163
	(.086)	(.076)	(.084)	
I worked with a group	3.509***	.118	-.195**	6960
	(.12)	(.106)	(.098)	
I solved problems on the blackboard	3.571***	.117	-.04	7040
	(.099)	(.075)	(.074)	
My teacher explained a topic	4.067***	.222***	.005	7093
	(.09)	(.064)	(.09)	
My teacher asked me to take mock exams	3.375***	.118	.027	7032
	(.109)	(.076)	(.069)	
My teacher graded my homework	4.281***	.131**	-.128*	7176
	(.07)	(.061)	(.073)	
Panel B. Spanish				
I used a textbook	2.775***	.301***	.177	7347
	(.108)	(.096)	(.124)	
My teacher assigned me homework	3.455***	.154*	.066	7336
	(.139)	(.085)	(.099)	
I copied from the blackboard	3.467***	.184**	-.005	7285
	(.114)	(.074)	(.084)	
I wrote something (e.g., a story)	2.88***	.072	-.029	7173
	(.112)	(.092)	(.075)	
I worked with a group	3.48***	.123	-.016	7143
	(.117)	(.094)	(.1)	
My teacher dictated a text to me	3.426***	.249***	.152	7126
	(.106)	(.072)	(.119)	
My teacher explained a topic	3.957***	.246***	.037	7176
	(.092)	(.066)	(.095)	
My teacher asked me to take mock exams	3.261***	.099	.052	7159
	(.111)	(.076)	(.082)	
My teacher graded my homework	4.27***	.097*	-.146**	7008
	(.078)	(.051)	(.064)	

Notes: (1) The table displays a different regression in every line, using two versions of equation (1): one that does not include school-by-grade level averages of test scores at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), and the number of non-missing observations (column 4). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 to 3 account for clustering at the school level. (4) All estimations include randomization strata fixed effects.

Table 9: ITT effects on student-reported interaction with teachers (2015)

	(1) Constant	(2) T1	(3) T2	(4) N
My teacher is nice when I ask for help	3.978*** (.086)	.212*** (.065)	-.003 (.072)	7468
My teacher gives me time to explain my ideas	3.901*** (.065)	.213*** (.058)	-.02 (.068)	7385
My teacher does not waste time	3.273*** (.088)	.127* (.073)	0 (.082)	7198
We know what we are doing and learning	3.66*** (.08)	.165** (.067)	.064 (.063)	7252
My teacher knows who understands the material	3.915*** (.071)	.209*** (.057)	.082 (.057)	7204
My teacher explains things in multiple ways	3.998*** (.078)	.166** (.076)	.027 (.069)	7345
My teacher asks us to reflect on what we read	4.015*** (.066)	.201*** (.068)	.018 (.072)	7217
My teacher makes us all try our best	4.048*** (.087)	.171** (.07)	-.003 (.064)	7319
Our schoolwork is interesting	4.144*** (.059)	.15*** (.047)	.001 (.056)	7275
Our homework helps us learn	4.434*** (.048)	.136*** (.045)	.036 (.05)	7295
My teacher checks that we understand	4.22*** (.069)	.164*** (.055)	-.007 (.065)	7162
My teacher asks us to explain our answers	3.883*** (.075)	.118 (.077)	.004 (.07)	7146
My teacher reviews what he/she teaches	4.098*** (.064)	.212*** (.057)	.051 (.053)	7179
My teacher writes comments when he/she grades	3.894*** (.073)	.159** (.077)	-.082 (.087)	7253

Notes: (1) The table displays a different regression in every line, using two versions of equation (1): one that does not include school-by-grade level averages of test scores at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), and the number of non-missing observations (column 4). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 to 3 account for clustering at the school level. (4) All estimations include randomization strata fixed effects.

Table 10: TOT effects on student achievement data (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		First stage			Second stage			
	Constant	T1	T2	Constant	T1	Act.	N	Controls?
Panel A. Grade 3								
Math (scaled)	.016	.84***	4.833***	.029	.315*	.032	3882	N
	(.179)	(.172)	(.315)	(.17)	(.118)	(.033)		
	.081	.873***	4.864***	.022	.312**	.032	3882	Y
	(.171)	(.176)	(.323)	(.173)	(.12)	(.034)		
Reading (scaled)	.038	.828***	4.819***	-.075	.338**	.022	3993	N
	(.175)	(.174)	(.292)	(.143)	(.107)	(.028)		
	.098	.861***	4.849***	-.047	.351**	.025	3993	Y
	(.17)	(.177)	(.3)	(.146)	(.108)	(.029)		
Panel B. Grade 5								
Math (scaled)	.123	.866***	4.757***	.185	.245	.045	4150	N
	(.186)	(.168)	(.3)	(.144)	(.118)	(.028)		
	.176	.894***	4.782***	.182	.243*	.045	4150	Y
	(.179)	(.172)	(.308)	(.143)	(.12)	(.029)		
Reading (scaled)	.14	.876***	4.748***	-.062	.343**	.04	4260	N
	(.187)	(.169)	(.301)	(.11)	(.116)	(.025)		
	.194	.904***	4.774***	-.066	.341**	.039	4260	Y
	(.18)	(.173)	(.308)	(.107)	(.118)	(.026)		

Notes: (1) The table displays a different regression in every line, using two versions of equation (4): one that does not include an index of school-level covariates from administrative data at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients from the first stage (columns 1-3), the coefficients from the second stage (columns 4-6), the number of non-missing observations (column 7), and whether the index of school-level covariates at baseline was included (column 8). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.

Appendix A Instruments

A.1 Baseline

A.1.1 Student assessments

The standardized tests of math and reading were designed by the *Centro de Estudios en Políticas Públicas* (CEPP) to assess what students ought to know and be able to do according to national and sub-national standards, including: (a) the *Contenidos Básicos Comunes*, the national curriculum; (b) the *Núcleos de Aprendizaje Prioritarios*, the contents in the national curriculum that the government identified as priorities; and (c) the *Diseño Curricular de La Rioja*, the province's curriculum.[32]

The tests assessed a wide array of domains and skills in math and reading at different difficulty levels. The math test included 30 multiple-choice items that covered number properties, geometry, measurement, probability and statistics. It assessed students' capacity to identify mathematical concepts, understand and use symbolic math, perform calculations using various strategies, and solve abstract and applied problems. The reading test included 30 multiple-choice items that featured informative and narrative texts. It assessed students' capacity to locate information in the text, understand the relationship between two parts of a text, identify the main idea of a text, and interpret the meaning of words from context. The item maps for both tests are available from the authors.

As Figure A.1 shows, our test design was successful in producing well-behaved distributions in math and reading, in grades 3 and 5, across all three years of the study. Only a few students were unable to answer any questions or all questions in each test correctly.

[Insert Figure A.1 here.]

Items were scored dichotomously and scaled using Item Response Theory (IRT). IRT models the relationship between a student's ability and the probability that he/she will answer a given item on a test correctly (Yen and Fitzpatrick 2006). IRT is used in LSAs for three main reasons. First, it allows each item to contribute differentially to the estimation of student ability (unlike percent-correct scores, which assign the same dichotomous score to each item). Second, it allows researchers to place different assessments on a common scale, provided that they share a subset of items and/or students. Third, it allows researchers to assess the performance of each individual item (which is particularly useful for test design).

[32] The assessments were developed by CEPP, the think tank that delivered the capacity-building workshops.

There are three IRT models that are frequently used (Harris 2005). We used a three-parameter logistic model, which estimates P_i, the probability that a student will answer item i correctly, based on: θ_p, student p's ability; a_i, the "discrimination" parameter (i.e., the slope of the Item Characteristic Curve (ICC) at the point of inflection, which reflects how well the item can differentiate between students of similar ability); b_i, the "difficulty" parameter (i.e., the point of inflection on the θ scale, which reflects where the item functions along the ability scale); and c_i, the "pseudo-guessing" parameter, the asymptotic probability that students will answer the item correctly by chance. The three-parameter model is thus given by:

$$P_i(\theta_p) = c_i + \frac{1 - c_i}{1 + e^{[-1.7a_i(\theta_p - b_i)]}} \tag{5}$$

where all parameters are defined defined as above. The model uses a logistic function to relate student ability and the item parameters to the probability of answering an item correctly.

We generated maximum likelihood estimates of student achievement, which are unbiased individual measures of ability, using the OpenIRT Stata program developed by Tristan Zajonc. Bayesian Markov chain Monte Carlo estimates are similar and available from the authors.

Figure A.2 shows the distribution of IRT scaled scores for math and reading, in grades 3 and 5, across all years of the study. Scaled scores were set to have a mean of 0 and a standard deviation of 1 within each subject and grade combination across all three years.

[Insert Figure A.2 here.]

A.1.2 Student survey

This survey included questions about students': (a) demographic characteristics; (b) home environment; (c) schooling trajectory; and (d) study habits.[33]

Table B.3 presents the results from this survey. As the table shows, about half of students in our sample were female. Half of them attended grade 3 and the other half grade 5. Most students had parents who had completed secondary school. The vast majority of students had basic household assets such as a fridge, a fan, a T.V., a washing machine, and a computer, but fewer of them had other assets such as a microwave, an air conditioner, or Internet, and less than a third of them had more than 20 books at home. More than a fifth of the students in our sample had previously repeated a grade. Finally, students received little academic support: 42% of them received homework everyday, but 41% reported doing homework alone (without the help of any relative), and only a marginal share of students received private tutoring. Students in T1 and T2 schools were comparable on nearly all the indicators mentioned above.

[33] This survey can be accessed at: `http://bit.ly/1qZeYHC`.

A.1.3 Teacher survey

This survey included questions about teachers': (a) demographic characteristics; (b) education and experience; (c) professional development; and (d) teaching practices.[34]

Table B.4 presents the results from this survey. As the table shows, about half of the teachers in our sample taught math and more than two thirds of them taught Spanish. Similarly, nearly half of them taught grade 3 and more than half taught grade 5. Most teachers (about two thirds) taught during the morning shift. Nearly all of them had a tertiary degree, which is the level at which teacher training often occurs. Access to professional development (PD) opportunities seems limited: only 42% of teachers reported attending more than two PD courses in 2013. Nearly half of teachers taught at other schools, but only 12% had another (non-teaching) job. Finally, only a minority of teachers had five or fewer years of teaching experience, but nearly half of them had spent five or fewer years at their current school, which suggests that teachers typically transfer across schools.

Teachers in T1 and T2 schools were comparable on most indicators, except for the share that teaches math and Spanish, and the share that teaches in the afternoon shift. Given that the lottery results had already been announced by the time that we collected these data, we cannot discard the possibility that these differences are attributable to treatment assignment.

A.2 Follow-ups

The survey of students included questions on aspects that we hypothesized may vary with time and treatment exposure, including: (a) study habits; and (b) the frequency of specific pedagogical practices of math and Spanish teachers (e.g., using a textbook or assigning homework).[35] Similarly, the survey of teachers focused on their: (a) education and experience; (b) initial training and professional development; (c) teaching practices; (d) monitoring and evaluation practices at their schools; (e) job satisfaction; and (f) perceptions of challenges that their schools face.[36] Finally, the survey of principals focused on their: (a) education and experience; (b) initial training and professional development; (c) school facilities and resources; (d) management practices; and (e) perceptions of challenges that their schools face.[37]

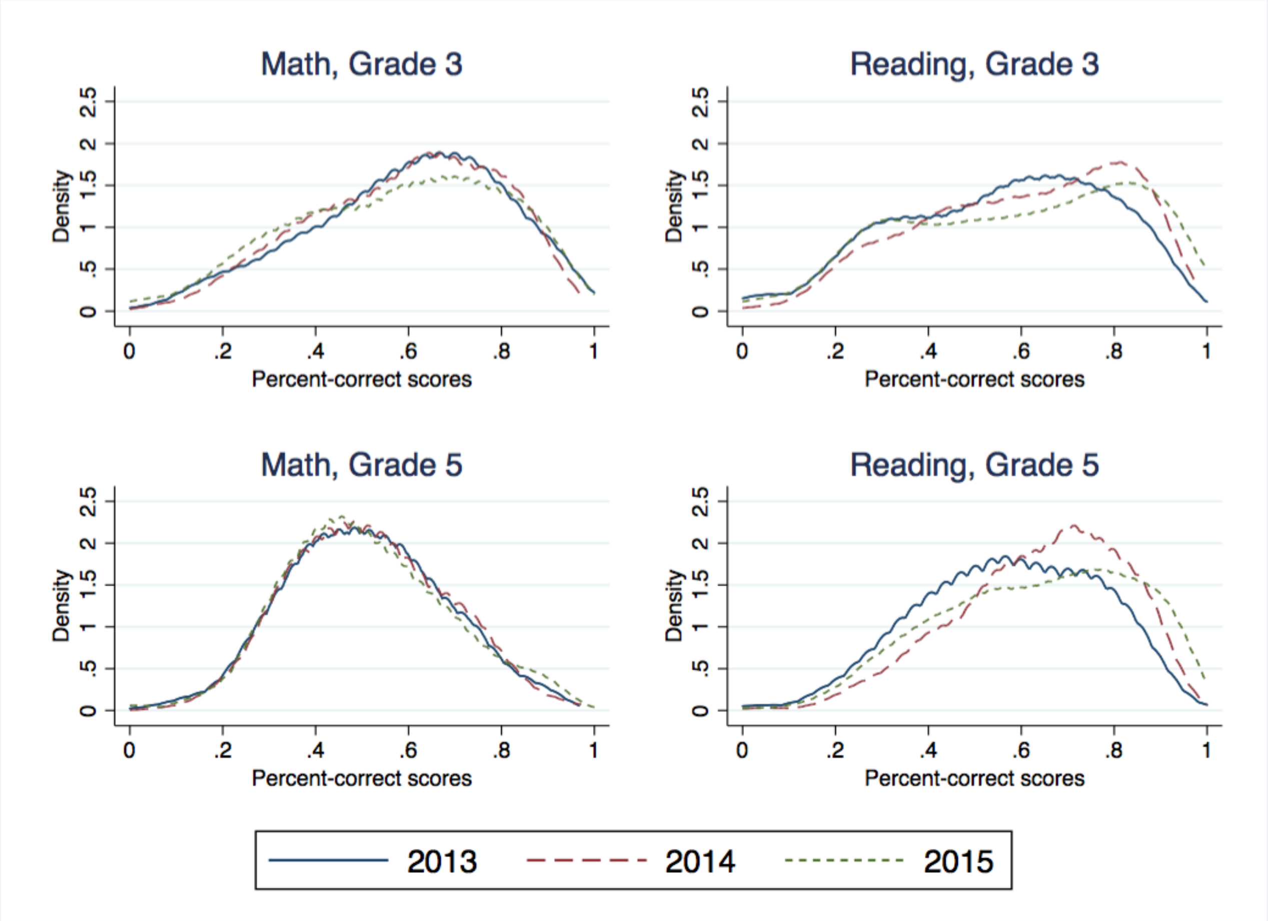
[34] This survey can be accessed at: `http://bit.ly/2OR1ni3`.

[35] The surveys can be accessed at: `http://bit.ly/1TOwuMt` (2014) and `http://bit.ly/1VrPBek` (2015).

[36] The surveys can be accessed at: `http://bit.ly/1sp2XNb` (2014) and `http://bit.ly/1THNgr0` (2015).

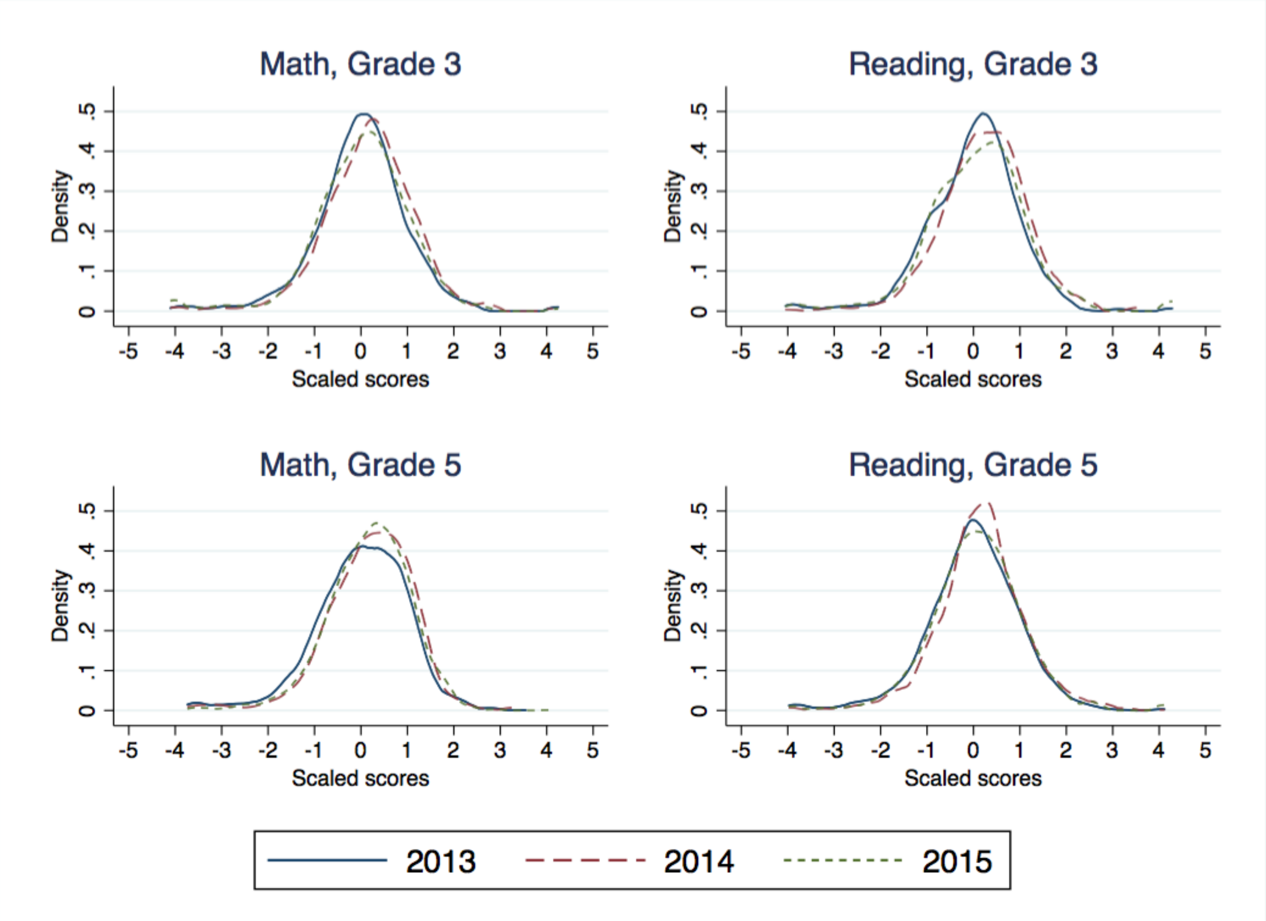
[37] The surveys can be accessed at: `http://bit.ly/1TTkCp6` (2014) and `http://bit.ly/1TUkwyO` (2015).

Figure A.1: Distribution of math and reading percent-correct scores (2013-2015)



Notes: This figure shows the distribution of percent-correct scores for each subject (math and reading) and year (2013-2015) in the impact evaluation.

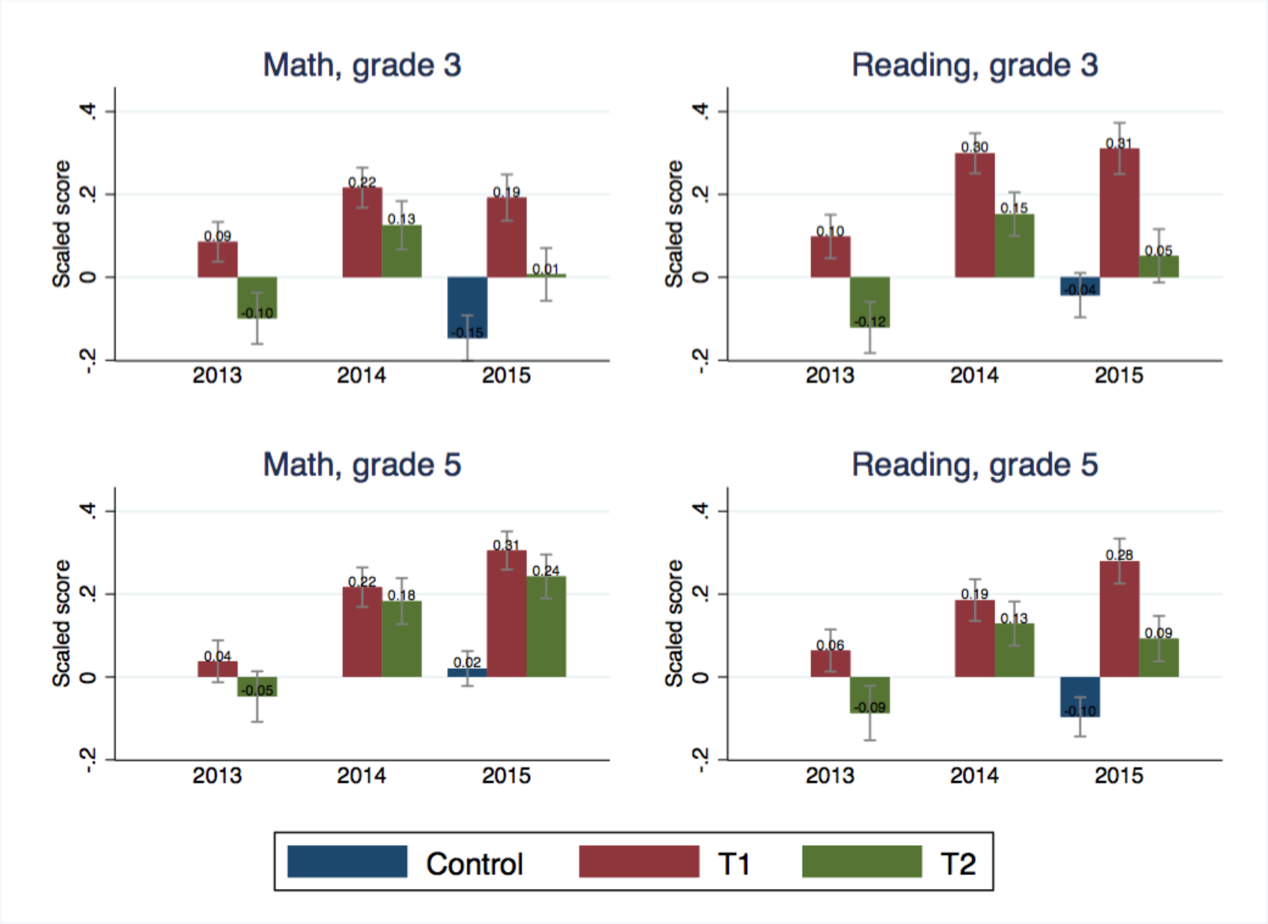
Figure A.2: Distribution of math and reading scaled scores (2013-2015)



Notes: This figure shows the distribution of scaled scores for each subject (math and reading) and year (2013-2015) in the impact evaluation. Scaled scores for each year were standardized using the pooled mean and standard deviation for T1 and T2 schools in 2013.

Appendix B Additional figures and tables

Figure B.1: IRT scaled mean scores, by experimental group (2013-2015)



Notes: The bars show the average IRT scaled scores for each subject (math and reading), grade (3 and 5), and year (2013-2015) in the impact evaluation. The whiskers show the standard deviation for each subject, grade, and year combination. Scaled scores for each year were standardized using the pooled mean and standard deviation for T1 and T2 schools in 2013.

Table B.1: Comparison of RCT and non-RCT schools on administrative data (2013)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Non-RCT schools					
Variable		Group 1:	Group 2:				
	All	Urban, semi-	Urban and	RCT	RCT-Group 1	RCT-Group 2	
	schools	urban, and rural	semi-urban	schools	schools	schools	N
Enrollment - Total	128.333	52.457	265.951	332.615	280.158***	66.664*	384
	(208.352)	(110.582)	(168.858)	(265.535)	(26.809)	(36.997)	
Enrollment - Grade 3	18.651	7.636	38.878	48.308	40.672***	9.43*	384
	(30.692)	(16.209)	(24.762)	(39.663)	(4)	(5.476)	
Enrollment - Grade 5	18.086	7.168	35.805	47.481	40.313***	11.676**	384
	(30.296)	(15.25)	(24.232)	(39.83)	(4.002)	(5.43)	
Urban school	.25	.118	.805	.606	.488***	-.199**	384
	(.434)	(.323)	(.401)	(.491)	(.052)	(.079)	
Semi-urban school	.127	.029	.195	.39	.362***	.195**	385
	(.334)	(.167)	(.401)	(.49)	(.049)	(.079)	
School has principal and vice-principal	.312	.136	.878	.781	.645***	-.097	385
	(.464)	(.343)	(.331)	(.416)	(.045)	(.066)	
School only has principal	.044	.021	.049	.105	.083***	.056	385
	(.206)	(.145)	(.218)	(.308)	(.031)	(.045)	

Notes: (1) The table shows the mean and standard deviations of all schools in the sample (column 1), C schools (column 2), T1 schools (column 3), and T2 schools (column 4). It also tests for differences across C and T1 schools (column 5), C and T2 schools (column 6), T1 and T2 schools (column 7), and shows the number of non-missing observations (column 8). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 5-7 account for clustering at the school level. (4) All specifications include randomization strata fixed effects.

Table B.2: Balancing checks on student achievement data (2013)

	(1)	(2)	(3)	(4)	(5)
Variable	T1 and T2	T1	T2	T2-T1	N
Panel A. Grade 3					
Math (scaled)	0	.085	-.099	-.185	2571
	(1)	(.907)	(1.09)	(.138)	
Reading (scaled)	0	.098	-.12	-.212	2351
	(1)	(.967)	(1.026)	(.143)	
F-statistic				.502	
p-value				.734	
Panel B. Grade 5					
Math (scaled)	0	.038	-.047	-.096	2519
	(1)	(.965)	(1.041)	(.141)	
Reading (scaled)	0	.064	-.087	-.156	2341
	(1)	(.953)	(1.055)	(.132)	
F-statistic				.485	
p-value				.746	

Notes: (1) The table shows the mean and standard deviations of all T1 and T2 schools (column 1), T1 schools (column 2), and T2 schools (column 3). It also tests for differences across T1 and T2 schools (column 4) and shows the number of non-missing observations (column 5). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in column 4 account for clustering at the school level. (4) The estimation for column 4 includes randomization strata fixed effects.

Table B.3: Balancing checks on student survey (2013)

Variable	(1) T1 and T2	(2) T1	(3) T2	(4) T2-T1	(5) N
Female	.498	.495	.503	.009	5238
	(.5)	(.5)	(.5)	(.013)	
Attends grade 3	.506	.495	.52	.029	5238
	(.5)	(.5)	(.5)	(.019)	
Attends grade 5	.494	.505	.48	-.029	5238
	(.5)	(.5)	(.5)	(.019)	
Father completed secondary school	.685	.697	.669	-.027	4577
	(.465)	(.46)	(.471)	(.037)	
Mother completed secondary school	.732	.751	.709	-.039	4667
	(.443)	(.433)	(.454)	(.035)	
Has fridge	.958	.957	.96	.007	4753
	(.2)	(.204)	(.195)	(.009)	
Has microwave	.533	.541	.524	-.009	4707
	(.499)	(.498)	(.5)	(.039)	
Has fan	.885	.892	.877	-.014	4837
	(.318)	(.31)	(.329)	(.014)	
Has air conditioner	.624	.628	.618	.003	4808
	(.484)	(.483)	(.486)	(.051)	
Has television	.986	.983	.989	.006	4913
	(.118)	(.128)	(.104)	(.004)	
Has washing machine	.953	.949	.959	.011	4803
	(.211)	(.22)	(.198)	(.01)	
Has computer	.825	.832	.816	-.013	4747
	(.38)	(.374)	(.388)	(.023)	
Has Internet	.656	.667	.642	-.018	5016
	(.475)	(.471)	(.48)	(.033)	
Has more than 20 books at home	.358	.379	.334	-.046	4711
	(.48)	(.485)	(.472)	(.031)	
Repeated a grade	.21	.183	.242	.063*	5238
	(.407)	(.387)	(.428)	(.037)	
Receives homework everyday	.456	.443	.472	.036	4825
	(.498)	(.497)	(.499)	(.039)	
Does homework alone	.444	.434	.456	.022	4834
	(.497)	(.496)	(.498)	(.023)	
Has private tutor	.05	.057	.042	-.013	4834
	(.219)	(.231)	(.202)	(.009)	
F-statistic				2.186	
p-value				.012	

Notes: (1) The table shows the mean and standard deviations of all T1 and T2 schools (column 1), T1 schools (column 2), and T2 schools (column 3). It also tests for differences across T1 and T2 schools (column 4) and shows the number of non-missing observations (column 5). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in column 4 account for clustering at the school level. (4) The estimation for column 4 includes randomization strata fixed effects.

Table B.4: Balancing checks on teacher survey (2013)

Variable	(1) T1 and T2	(2) T1	(3) T2	(4) T2-T1	(5) N
Teaches math	.489	.444	.543	.096*	307
	(.501)	(.498)	(.5)	(.054)	
Teaches Spanish	.7	.657	.754	.088*	307
	(.459)	(.476)	(.432)	(.044)	
Teaches grade 3	.472	.482	.46	-.022	307
	(.5)	(.501)	(.5)	(.05)	
Teaches grade 5	.524	.518	.533	.015	307
	(.5)	(.501)	(.501)	(.049)	
Teaches in the morning shift	.66	.608	.725	.092	309
	(.474)	(.49)	(.448)	(.089)	
Teaches in the afternoon shift	.424	.509	.319	-.192**	309
	(.495)	(.501)	(.468)	(.092)	
Has a tertiary degree	.928	.935	.921	-.019	307
	(.258)	(.248)	(.271)	(.037)	
Took more than 2 PD courses this year	.416	.45	.374	-.086	310
	(.494)	(.499)	(.486)	(.062)	
Teaches at multiple schools	.5	.474	.533	.089	308
	(.501)	(.501)	(.501)	(.084)	
Has another job	.118	.118	.118	.002	305
	(.323)	(.324)	(.323)	(.04)	
Has 5 or fewer years of experience	.123	.147	.094	-.052	309
	(.329)	(.355)	(.292)	(.036)	
Has 5 or fewer years of experience at this school	.5	.515	.482	-.016	308
	(.501)	(.501)	(.502)	(.074)	
F-statistic				11.138	
p-value				0	

Notes: (1) The table shows the mean and standard deviations of all T1 and T2 schools (column 1), T1 schools (column 2), and T2 schools (column 3). It also tests for differences across T1 and T2 schools (column 4) and shows the number of non-missing observations (column 5). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in column 4 account for clustering at the school level. (4) The estimation for column 4 includes randomization strata fixed effects.

Table B.5: Treatment dosage (2015)

Variable	(1) C	(2) T1	(3) T2	(4) T1-C	(5) T2-C	(6) T2-T1	(7) N
Panel A. Principal survey							
Students took standardized tests	.389	.962	.931	.57***	.539***	-.033	90
	(.494)	(.196)	(.258)	(.092)	(.098)	(.063)	
Compared test results w/those of province	.241	.679	.75	.438***	.5***	.066	84
	(.435)	(.476)	(.441)	(.12)	(.121)	(.124)	
Compared test results w/those of other schools	.241	.741	.519	.491***	.3**	-.177	82
	(.435)	(.447)	(.509)	(.119)	(.13)	(.133)	
Tracked test results over time	.688	1	.966	.312***	.281***	-.039	87
	(.471)	(0)	(.186)	(.084)	(.094)	(.039)	
Panel B. Implementation monitoring							
Number of reports received	1	3	3	2	2	0	103
	(0)	(0)	(0)	(0)	(0)	(0)	
Number of workshops attended	0	.233	3.4	.228***	3.38***	3.156***	103
	(0)	(.43)	(.855)	(.078)	(.161)	(.181)	
Number of visits received	0	.733	1.467	.749***	1.449***	.688***	103
	(0)	(.583)	(.571)	(.099)	(.107)	(.141)	

Notes: (1) The table shows the mean and standard deviations of control schools (column 1), T1 schools (column 2), and T2 schools (column 3). It also tests for differences across C and T1 schools (column 4), C and T2 schools (column 5), T1 and T2 schools (column 6), and shows the number of non-missing observations (column 7). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 5-7 account for clustering at the school level. (4) All specifications include randomization strata fixed effects.

Table B.6: ITT effects on student achievement data (2015)

	(1) Constant	(2) T2	(3) N	(4) Controls?
Panel A. Grade 3				
Math (scaled)	.306	-.178	2369	N
	(.204)	(.167)		
	.19	.002	2369	Y
	(.177)	(.105)		
Reading (scaled)	-.163	-.233	2311	N
	(.176)	(.149)		
	-.25	-.078	2311	Y
	(.191)	(.095)		
Panel B. Grade 5				
Math (scaled)	.491***	-.07	2489	N
	(.18)	(.145)		
	.38**	-.03	2489	Y
	(.173)	(.099)		
Reading (scaled)	.142	-.208	2422	N
	(.138)	(.144)		
	.262*	-.107	2422	Y
	(.149)	(.096)		

Notes: (1) The table displays a different regression in every line, using two versions of equation (1): one that does not include school-by-grade level averages of test scores at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and T2 dummy (column 2), the number of non-missing observations (column 3), and whether the school-by-grade level average of test scores at baseline was included (column 4). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 account for clustering at the school level. (4) All estimations include randomization strata fixed effects.

Table B.7: ITT effects on math achievement, by content domain (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
					F-tests	
	Constant	T1	T2	N	$\beta_1 = \beta_2 = 0$	$\beta_1 = \beta_2$
Panel A. Grade 3						
Numbers (percent-correct)	.535***	.067**	.039	4439	2.984	.544
	(.035)	(.028)	(.035)		(.055)	(.463)
Geometry (percent-correct)	.555***	.068**	.035	4439	3.387	.89
	(.036)	(.027)	(.031)		(.038)	(.348)
Measurement (percent-correct)	.459***	.087***	.047	4439	5.344	1.013
	(.037)	(.027)	(.035)		(.006)	(.316)
Panel B. Grade 5						
Numbers (percent-correct)	.472***	.043	.042	4664	1.514	.001
	(.032)	(.03)	(.032)		(.225)	(.976)
Geometry (percent-correct)	.456***	.073***	.035	4664	4.045	1.531
	(.022)	(.027)	(.024)		(.02)	(.219)
Measurement (percent-correct)	.396***	.07**	.035	4664	3.3	1.216
	(.03)	(.029)	(.023)		(.041)	(.273)
Statistics (percent-correct)	.633***	.055*	.022	4664	1.832	.956
	(.027)	(.029)	(.028)		(.165)	(.331)

Notes: (1) The table displays a different regression in every line, using two versions of equation (2): one that does not include an index of school-level covariates from administrative data at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), and the number of non-missing observations (column 4). It also shows the results from two F-tests: one testing whether the coefficients on the T1 and T2 dummies were jointly statistically significant (column 5) and another one testing whether the coefficient on the T1 dummy is statistically significantly different from the coefficient on the T2 dummy (column 6). In both columns, the F-statistic is shown with its associated p-value (in parentheses). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.

Table B.8: ITT effects on reading achievement, by content domain (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
					F-tests	
	Constant	T1	T2	N	$\beta_1 = \beta_2 = 0$	$\beta_1 = \beta_2$
Panel A. Grade 3						
Narrative texts (percent-correct)	.541***	.081***	.021	4439	3.882	2.297
	(.035)	(.029)	(.036)		(.024)	(.133)
Informative texts (percent-correct)	.547***	.07**	.024	4439	3.244	1.269
	(.031)	(.027)	(.036)		(.043)	(.263)
Short texts (percent-correct)	.548***	.093***	.01	4439	4.882	3.431
	(.036)	(.03)	(.042)		(.009)	(.067)
Panel B. Grade 5						
Narrative texts (percent-correct)	.593***	.068***	.019	4664	3.925	2.216
	(.023)	(.024)	(.028)		(.023)	(.14)
Informative texts (percent-correct)	.581***	.073***	.029	4664	5.077	1.697
	(.025)	(.023)	(.029)		(.008)	(.196)
Short texts (percent-correct)	.65***	.059**	.001	4664	3.382	3.277
	(.025)	(.024)	(.03)		(.038)	(.073)

Notes: (1) The table displays a different regression in every line, using two versions of equation (2): one that does not include an index of school-level covariates from administrative data at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), and the number of non-missing observations (column 4). It also shows the results from two F-tests: one testing whether the coefficients on the T1 and T2 dummies were jointly statistically significant (column 5) and another one testing whether the coefficient on the T1 dummy is statistically significantly different from the coefficient on the T2 dummy (column 6). In both columns, the F-statistic is shown with its associated p-value (in parentheses). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.

Table B.9: ITT effects on math achievement, by cognitive domain (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
					F-tests	
	Constant	T1	T2	N	$\beta_1 = \beta_2 = 0$	$\beta_1 = \beta_2$
Panel A. Grade 3						
Communicating (percent-correct)	.534***	.058**	.044	4439	2.932	.191
	(.033)	(.026)	(.029)		(.058)	(.663)
Knowing (percent-correct)	.537***	.078***	.042	4439	4.693	.97
	(.036)	(.026)	(.033)		(.011)	(.327)
Algorithms (percent-correct)	.556***	.071**	.041	4439	2.878	.528
	(.037)	(.03)	(.038)		(.061)	(.469)
Problems (percent-correct)	.45***	.071**	.035	4439	3.168	.864
	(.036)	(.028)	(.035)		(.046)	(.355)
Panel B. Grade 5						
Communicating (percent-correct)	.629***	.054*	.025	4664	1.796	.688
	(.028)	(.029)	(.029)		(.171)	(.409)
Knowing (percent-correct)	.531***	.071***	.033	4664	4.006	1.628
	(.024)	(.026)	(.024)		(.021)	(.205)
Algorithms (percent-correct)	.526***	.031	.029	4664	.519	.003
	(.037)	(.035)	(.039)		(.596)	(.958)
Problems (percent-correct)	.328***	.062**	.05*	4664	3.238	.113
	(.032)	(.029)	(.026)		(.043)	(.737)

Notes: (1) The table displays a different regression in every line, using two versions of equation (2): one that does not include an index of school-level covariates from administrative data at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), and the number of non-missing observations (column 4). It also shows the results from two F-tests: one testing whether the coefficients on the T1 and T2 dummies were jointly statistically significant (column 5) and another one testing whether the coefficient on the T1 dummy is statistically significantly different from the coefficient on the T2 dummy (column 6). In both columns, the F-statistic is shown with its associated p-value (in parentheses). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.

Table B.10: ITT effects on reading achievement, by cognitive domain (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
					F-tests	
	Constant	T1	T2	N	$\beta_1 = \beta_2 = 0$	$\beta_1 = \beta_2$
Panel A. Grade 3						
Explicit info. (percent-correct)	.581***	.076***	.022	4439	3.963	1.872
	(.033)	(.027)	(.037)		(.022)	(.174)
Implicit info. (percent-correct)	.571***	.077***	.016	4439	3.631	2.255
	(.033)	(.029)	(.037)		(.03)	(.136)
Analyzing texts (percent-correct)	.475***	.082***	.024	4439	3.933	1.985
	(.032)	(.029)	(.036)		(.023)	(.162)
Reflecting (percent-correct)	.53***	.075**	.029	4439	2.748	1.306
	(.039)	(.032)	(.036)		(.069)	(.256)
Panel B. Grade 5						
Explicit info. (percent-correct)	.662***	.067***	.015	4664	4.48	2.739
	(.023)	(.022)	(.028)		(.014)	(.101)
Implicit info. (percent-correct)	.558***	.059**	.013	4664	3.102	1.966
	(.024)	(.024)	(.029)		(.049)	(.164)
Analyzing texts (percent-correct)	.57***	.081***	.034	4664	4.911	1.724
	(.025)	(.026)	(.03)		(.009)	(.192)
Reflecting (percent-correct)	.562***	.067**	.02	4664	2.32	1.508
	(.031)	(.031)	(.032)		(.103)	(.222)

Notes: (1) The table displays a different regression in every line, using two versions of equation (2): one that does not include an index of school-level covariates from administrative data at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), and the number of non-missing observations (column 4). It also shows the results from two F-tests: one testing whether the coefficients on the T1 and T2 dummies were jointly statistically significant (column 5) and another one testing whether the coefficient on the T1 dummy is statistically significantly different from the coefficient on the T2 dummy (column 6). In both columns, the F-statistic is shown with its associated p-value (in parentheses). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.

Table B.11: ITT effects on student-reported teacher attendance and punctuality (2015)

	(1) Constant	(2) T1	(3) T2	(4) N
Freq. of teacher was absent to school	2.835***	-.101	.043	7308
	(.138)	(.112)	(.115)	
Freq. of teacher arrived late to school	1.827***	-.062	.063	7363
	(.092)	(.085)	(.087)	
Freq. of teacher started class late	2.078***	-.112	.005	7158
	(.096)	(.08)	(.068)	
Freq. of teacher ended class late	2.176***	-.107	.041	7138
	(.103)	(.065)	(.068)	
Freq. of teacher started class early	1.784***	-.237***	-.035	7171
	(.101)	(.062)	(.075)	
Freq. of teacher ended class early	2.343***	-.195**	-.082	7098
	(.118)	(.084)	(.078)	
Freq. of left school early	2.224***	-.159**	-.077	7274
	(.106)	(.08)	(.084)	

Notes: (1) The table displays a different regression in every line, using two versions of equation (1): one that does not include school-by-grade level averages of test scores at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), and the number of non-missing observations (column 4). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 to 3 account for clustering at the school level. (4) All estimations include randomization strata fixed effects.

Table B.12: ITT effects on student achievement, by total enrollment (2015)

	(1) Constant	(2) T1	(3) T2	(4) Enr.	(5) T1 × Enr.	(6) T2 × Enr.	(7) N
Panel A. Grade 3							
Math (scaled)	-.137	.227	.163	-.00001	.00022	-.00002	3882
	(.111)	(.283)	(.324)	(.00012)	(.00058)	(.00072)	
Reading (scaled)	-.13	.142	.042	.00013*	.00047	.00016	3993
	(.096)	(.223)	(.278)	(.00007)	(.00044)	(.00057)	
Panel B. Grade 5							
Math (scaled)	.013	.136	.493*	.00001	.00031	-.00057	4150
	(.091)	(.232)	(.249)	(.00007)	(.00054)	(.00048)	
Reading (scaled)	-.081	.174	.144	-.00002	.00041	.00008	4260
	(.079)	(.257)	(.24)	(.00006)	(.0006)	(.00056)	

Notes: (1) The table displays a different regression in every line, using a version of equation (2) that includes total enrollment at each school and its interactions with the treatment dummies. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1), the T1 and T2 dummies (columns 2 and 3), total enrollment (column 4), its interactions with the treatment dummies (columns 5 and 6), and the number of non-missing observations (column 7). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations exclude randomization strata fixed effects.

Table B.13: ITT effects on student achievement, by grade enrollment (2015)

	(1) Constant	(2) T1	(3) T2	(4) Enr.	(5) T1 × Enr.	(6) T2 × Enr.	(7) N
Panel A. Grade 3							
Math (scaled)	-.128	.228	.143	-.0002	.0014	.0001	3882
	(.112)	(.261)	(.274)	(.0009)	(.0036)	(.0039)	
Reading (scaled)	-.126	.145	.079	.0009	.0031	.0004	3993
	(.098)	(.211)	(.242)	(.0006)	(.003)	(.003)	
Panel B. Grade 5							
Math (scaled)	.014	.157	.536**	.0001	.0019	-.0047	4150
	(.087)	(.237)	(.258)	(.0004)	(.0037)	(.0035)	
Reading (scaled)	-.083	.168	.125	-.0001	.003	.0008	4260
	(.076)	(.257)	(.232)	(.0003)	(.0042)	(.0038)	

Notes: (1) The table displays a different regression in every line, using a version of equation (2) that includes grade-specific enrollment at each school and its interactions with the treatment dummies. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1), the T1 and T2 dummies (columns 2 and 3), total enrollment (column 4), its interactions with the treatment dummies (columns 5 and 6), and the number of non-missing observations (column 7). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations exclude randomization strata fixed effects.

Table B.14: ITT effects on student achievement, by school size (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
						T1 ×	T2 ×	T1 ×	T2 ×	
	Constant	T1	T2	Small	Med.	Small	Small	Med.	Med.	N
Panel A. Grade 3										
Math (scaled)	-.178**	.415***	.236	.315	.028	-.494	-.084	-.174	-.331	3904
	(.089)	(.149)	(.203)	(.287)	(.155)	(.415)	(.398)	(.248)	(.33)	
Reading (scaled)	-.015	.425***	.108	.236	-.222	-.642**	-.403	-.081	.138	4015
	(.077)	(.139)	(.162)	(.207)	(.158)	(.294)	(.332)	(.23)	(.317)	
Panel B. Grade 5										
Math (scaled)	.03	.344**	.13	.128	-.086	-.407	.418	-.115	.139	4170
	(.081)	(.156)	(.154)	(.203)	(.115)	(.272)	(.331)	(.216)	(.281)	
Reading (scaled)	-.109	.48***	.231	.084	.031	-.322	.039	-.33	-.217	4280
	(.067)	(.152)	(.152)	(.116)	(.125)	(.198)	(.27)	(.235)	(.231)	

Notes: (1) The table displays a different regression in every line, using a version of equation (2) that includes dummies for school size terciles (omitting large schools) and their interactions with the treatment dummies. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1), the T1 and T2 dummies (columns 2 and 3), the small and medium school dummies (columns 4 and 5), their interactions with the treatment dummies (columns 6 to 9), and the number of non-missing observations (column 10). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations exclude randomization strata fixed effects.

Table B.15: ITT effects on student achievement, by geographic location (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
					T1 ×	T2 ×	
	Constant	T1	T2	Urban	Urban	Urban	N
Panel A. Grade 3							
Math (scaled)	.206	.115	-.158	-.184	.277	.395	3882
	(.204)	(.24)	(.238)	(.195)	(.274)	(.303)	
Reading (scaled)	.107	.16	-.298	-.158	.24	.51**	3993
	(.18)	(.204)	(.205)	(.18)	(.236)	(.257)	
Panel B. Grade 5							
Math (scaled)	.201	.087	.266	.074	.218	-.072	4150
	(.159)	(.193)	(.236)	(.142)	(.233)	(.274)	
Reading (scaled)	.035	.285	-.011	-.085	.113	.247	4260
	(.146)	(.249)	(.208)	(.156)	(.278)	(.247)	

Notes: (1) The table displays a different regression in every line, using a version of equation (2) that includes a dummy for urban schools and its interactions with the treatment dummies. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1), the T1 and T2 dummies (columns 2 and 3), a dummy for urban schools (column 4), its interactions with treatment dummies (columns 5 and 6), and the number of non-missing observations (column 7). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.

Table B.16: ITT effects on student achievement, by item familiarity (2015)

	(1)	(2)	(3)	(4)	(5)	(6)
					F-tests	
	Constant	T1	T2	N	$\beta_1 = \beta_2 = 0$	$\beta_1 = \beta_2$
Panel A. Grade 3						
Math, familiar (percent-correct)	.542***	.062**	.027	4439	3.827	1
	(.035)	(.025)	(.032)		(.053)	(.32)
Math, non-familiar (percent-correct)	.513***	.076***	.046	4439	6.03	.623
	(.035)	(.028)	(.034)		(.016)	(.432)
Reading, familiar (percent-correct)	.556***	.084***	.018	4439	3.536	2.553
	(.032)	(.029)	(.039)		(.063)	(.113)
Reading, non-familiar (percent-correct)	.537***	.075***	.023	4439	3.838	1.722
	(.033)	(.028)	(.035)		(.053)	(.192)
Panel B. Grade 5						
Math, familiar (percent-correct)	.465***	.056*	.041	4664	4.632	.218
	(.031)	(.03)	(.026)		(.034)	(.642)
Math, non-familiar (percent-correct)	.477***	.059**	.036	4664	4.955	.493
	(.027)	(.027)	(.026)		(.028)	(.484)
Reading, familiar (percent-correct)	.641***	.055**	.003	4664	2.209	3.016
	(.024)	(.021)	(.027)		(.14)	(.085)
Reading, non-familiar (percent-correct)	.574***	.073***	.026	4664	5.344	1.819
	(.024)	(.026)	(.03)		(.023)	(.18)

Notes: (1) The table displays a different regression in every line, using two versions of equation (2): one that only includes items (i.e., included in previous assessment rounds) and one that only includes items (i.e., not included in previous assessment rounds). Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1) and on the T1 and T2 dummies (columns 2 and 3), and the number of non-missing observations (column 4). It also shows the results from two F-tests: one testing whether the coefficients on the T1 and T2 dummies were jointly statistically significant (column 5) and another one testing whether the coefficient on the T1 dummy is statistically significantly different from the coefficient on the T2 dummy (column 6). In both columns, the F-statistic is shown with its associated p-value (in parentheses). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.

Table B.17: ITT effects on student achievement, by absenteeism (2015)

	(1) Constant	(2) T1	(3) T2	(4) Abs.	(5) N
Panel A. Grade 3					
Math (scaled)	.03	.343***	.156		3882
	(.168)	(.119)	(.154)		
	.269	.316***	.124	-.013**	3882
	(.187)	(.113)	(.149)	(.006)	
Reading (scaled)	-.074	.356***	.108		3993
	(.141)	(.109)	(.131)		
	.239	.321***	.092	-.02***	3993
	(.146)	(.09)	(.109)	(.005)	
Panel B. Grade 5					
Math (scaled)	.191	.284**	.214*		4150
	(.139)	(.116)	(.124)		
	.352**	.256**	.225*	-.013**	4150
	(.165)	(.105)	(.119)	(.006)	
Reading (scaled)	-.056	.378***	.188*		4260
	(.105)	(.114)	(.112)		
	.079	.37***	.221**	-.015***	4260
	(.12)	(.111)	(.099)	(.005)	

Notes: (1) The table displays a different regression in every line, using two versions of equation (2): one that does not account for the share of absent students on testing day and one that does. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1), the T1 and T2 dummies (columns 2 and 3), the share of absent students (column 4), and the number of non-missing observations (column 5). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations exclude randomization strata fixed effects.

Table B.18: ITT effects on student achievement, by supervisor assignment (2015)

	(1) Constant	(2) T1	(3) T2	(4) Sup.	(5) T1 × Sup.	(6) N
Panel A. Grade 3						
Math (scaled)	-.009	.398	.138	.082	-.086	3882
	(.186)	(.312)	(.152)	(.126)	(.337)	
Reading (scaled)	-.102	.456*	.095	.057	-.126	3993
	(.16)	(.256)	(.132)	(.12)	(.277)	
Panel B. Grade 5						
Math (scaled)	.19	.193	.214*	-.001	.097	4150
	(.151)	(.281)	(.128)	(.098)	(.31)	
Reading (scaled)	-.103	.542	.165	.104	-.21	4260
	(.117)	(.416)	(.116)	(.087)	(.432)	

Notes: (1) The table displays a different regression in every line, using a version of equation (2) that includes a dummy for supervisors who are responsible for control and T1 schools, and the interaction between this dummy and the T1 dummy. Each line shows the dependent variable of the regression on the left, the coefficients on the constant term (column 1), the T1 and T2 dummies (columns 2 and 3), the supervisor of control and T1 schools dummy (column 4), the interaction between that dummy and the T1 dummy (column 5), and the number of non-missing observations (column 6). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations exclude randomization strata fixed effects.

Table B.19: Dose-response relationship between workshops and student achievement (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		First stage			Second stage			
	Constant	T1	T2	Constant	T1	Work.	N	Controls?
Panel A. Grade 3								
Math (scaled)	-.115	.286***	3.474***	.035	.33**	.045	3882	N
	(.144)	(.106)	(.221)	(.168)	(.118)	(.046)		
	-.06	.314***	3.499***	.027	.326**	.044	3882	Y
	(.142)	(.109)	(.225)	(.171)	(.121)	(.047)		
Reading (scaled)	-.103	.278***	3.461***	-.071	.348**	.031	3993	N
	(.143)	(.104)	(.207)	(.14)	(.107)	(.039)		
	-.052	.305***	3.486***	-.043	.361**	.035	3993	Y
	(.142)	(.107)	(.211)	(.143)	(.108)	(.04)		
Panel B. Grade 5								
Math (scaled)	-.049	.283***	3.407***	.194	.266*	.063	4150	N
	(.141)	(.103)	(.207)	(.143)	(.117)	(.039)		
	-.005	.305***	3.429***	.19	.264*	.062	4150	Y
	(.139)	(.106)	(.213)	(.141)	(.12)	(.04)		
Reading (scaled)	-.033	.285***	3.408***	-.054	.362**	.055	4260	N
	(.143)	(.104)	(.21)	(.109)	(.115)	(.035)		
	.012	.309***	3.43***	-.059	.36**	.055	4260	Y
	(.14)	(.107)	(.214)	(.106)	(.118)	(.036)		

Notes: (1) The table displays a different regression in every line, using two versions of equation (4): one that does not include an index of school-level covariates from administrative data at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients from the first stage (columns 1-3), the coefficients from the second stage (columns 4-6), the number of non-missing observations (column 7), and whether the index of school-level covariates at baseline was included (column 8). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.

Table B.20: Dose-response relationship between visits and student achievement (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	First stage			Second stage				
	Constant	T1	T2	Constant	T1	Visits	N	Controls?
Panel A. Grade 3								
Math (scaled)	.131	.554***	1.36***	.015	.279*	.115	3882	N
	(.123)	(.138)	(.123)	(.179)	(.126)	(.12)		
	.141	.56***	1.364***	.008	.277*	.113	3882	Y
	(.12)	(.138)	(.126)	(.18)	(.126)	(.124)		
Reading (scaled)	.141	.551***	1.358***	-.085	.313**	.079	3993	N
	(.119)	(.134)	(.119)	(.151)	(.112)	(.101)		
	.15	.556***	1.363***	-.058	.322**	.09	3993	Y
	(.117)	(.134)	(.122)	(.152)	(.113)	(.104)		
Panel B. Grade 5								
Math (scaled)	.172	.584***	1.35***	.163	.191	.159	4150	N
	(.115)	(.141)	(.117)	(.15)	(.129)	(.101)		
	.181	.588***	1.354***	.161	.19	.158	4150	Y
	(.113)	(.142)	(.119)	(.148)	(.129)	(.104)		
Reading (scaled)	.173	.591***	1.34***	-.081	.295*	.141	4260	N
	(.115)	(.142)	(.117)	(.116)	(.124)	(.092)		
	.182	.595***	1.344***	-.084	.294*	.139	4260	Y
	(.113)	(.142)	(.119)	(.112)	(.125)	(.095)		

Notes: (1) The table displays a different regression in every line, using two versions of equation (4): one that does not include an index of school-level covariates from administrative data at baseline and one that does. Each line shows the dependent variable of the regression on the left, the coefficients from the first stage (columns 1-3), the coefficients from the second stage (columns 4-6), the number of non-missing observations (column 7), and whether the index of school-level covariates at baseline was included (column 8). (2) * significant at 10%; ** significant at 5%; *** significant at 1%. (3) Standard errors in columns 1 and 2 are clustered at the school level. (4) All estimations include randomization strata fixed effects.